Universidade Federal da Paraíba
Centro de Informática
Mestrado em Informática

Uma Arquitetura de Referência para Explicabilidade como Serviço na Saúde H-XAIaaS: *Health - eXplainable Artificial Intelligence as a Service*

Thiago Cunha Montenegro

João Pessoa - PB
Agosto de 2025

Thiago Cunha Montenegro

Uma Arquitetura de Referência para Explicabilidade como Serviço na Saúde H-XAIaaS: *Health - eXplainable Artificial Intelligence as a Service*

Dissertação de Mestrado em Informática do Centro de Informática da Universidade Federal da Paraíba (UFPB), como requisito para obtenção do grau de Mestre em em Informática.

Orientadora: Profa. Dra. Natasha Correia Queiroz Lino

João Pessoa - PB
Agosto de 2025

Catalogação na publicação
Seção de Catalogação e Classificação

M777a Montenegro, Thiago Cunha.
Uma arquitetura de referência para explicabilidade como serviço na saúde H-XAIaaS : Health - eXplainable Artificial Intelligence as a Service / Thiago Cunha Montenegro. - João Pessoa, 2025.
113 f. : il.

Orientação: Natasha Correia Queiroz Lino.
Dissertação (Mestrado) - UFPB/CI.

1. Inteligência artificial explicável. 2. Aprendizagem de máquina. 3. Suporte à decisão clínica - Sistemas. 4. Arquitetura de referência. 5. XAI. 6. Sistemas explicáveis na saúde. I. Lino, Natasha Correia Queiroz. II. Título.

UFPB/BC CDU 004.8(043)

Elaborado por Larissa Silva Oliveira de Mesquita - CRB-15/746

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de Thiago Cunha Montenegro, candidato ao título de Mestre em Informática na área de Sistemas de Computação, realizada em 30 de agosto de 2024.

Aos trinta dias do mês de agosto do ano de dois mil e vinte e quatro, às nove horas, no Centro de Informática da Universidade Federal da Paraíba, reuniram-se os membros da Banca Examinadora constituída para julgar o Trabalho Final do discente Thiago Cunha Montenegro, vinculado a esta Universidade sob a matrícula nº 20221004958, candidato ao grau de Mestre em Informática, na área de *"Sistemas de Computação"*, na linha de pesquisa *"Computação Distribuída"*, do Programa de Pós-Graduação em Informática. A comissão examinadora foi composta pelos professores: Natasha Correia Queiroz Lino, Orientadora e Presidente da banca; Clauirton de Albuquerque Siebra, Examinador Interno; Fernando José Ribeiro Sales, Examinador Externo à Instituição. Dando início aos trabalhos, a Presidente da Banca cumprimentou os presentes, comunicou a finalidade da reunião e passou a palavra ao candidato para que ele fizesse a exposição oral do trabalho de dissertação intitulado "**Uma Arquitetura de Referência para Explicabilidade como Serviço na Saúde - H-XAIaaS (Health - eXplainable Artificial Intelligence as a Service)**". Concluída a exposição, o candidato foi arguido pela Banca Examinadora que emitiu o seguinte parecer: "***aprovado***". Do ocorrido, eu, Gilberto Farias de Sousa Filho, Coordenador do Programa de Pós-Graduação em Informática, lavrei a presente ata que vai assinada por mim e pelos membros da Banca Examinadora. João Pessoa, 30 de agosto de 2024.

gov.br
Documento assinado digitalmente
GILBERTO FARIAS DE SOUSA FILHO
Data: 05/09/2024 10:34:15-0300
Verifique em https://validar.iti.gov.br

Gilberto Farias de Sousa Filho
Coordenador do Programa de Pós-Graduação em Informática

gov.br
Documento assinado digitalmente
NATASHA CORREIA QUEIROZ LINO
Data: 05/09/2024 09:30:41-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Natasha Correia Queiroz Lino
Orientadora (PPGI-UFPB)

gov.br
Documento assinado digitalmente
CLAUIRTON DE ALBUQUERQUE SIEBRA
Data: 31/08/2024 09:18:10-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Clauirton de Albuquerque Siebra
Examinador Interno (PPGI-UFPB)

gov.br
Documento assinado digitalmente
FERNANDO JOSE RIBEIRO SALES
Data: 30/08/2024 17:40:38-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Fernando José Ribeiro Sales
Examinador Externo à Instituição (UFPE)

Thiago Cunha Montenegro

Uma Arquitetura de Referência para Explicabilidade como Serviço na Saúde H-XAIaaS: *Health - eXplainable Artificial Intelligence as a Service*

Dissertação de Mestrado em Informática do Centro de Informática da Universidade Federal da Paraíba (UFPB), como requisito para obtenção do grau de Mestre em em Informática.

Trabalho aprovado. João Pessoa - PB, 30 de agosto de 2024:

Profa. Dra. Natasha Correia Queiroz Lino
Orientadora

Prof. Dr. Clauirton de Albuquerque Siebra
Examinador Interno

Prof. Dr. Fernando José Ribeiro Sales
Examinador Externo

João Pessoa - PB
Agosto de 2025

Agradecimentos

Em primeiro lugar, gostaria de expressar minha profunda gratidão à Professora Natasha por seus ensinamentos, orientação e constante suporte ao longo de todo o mestrado. Agradeço também aos meus pais, Rozevania e Julierme, pelo apoio incondicional e incentivo contínuo. À minha noiva, Rayla, sou imensamente grato por seu companheirismo, carinho e compreensão em todos os momentos desta jornada. Agradeço ainda aos meus colegas, com quem compartilhei vitórias e desafios ao longo da experiência do mestrado. Por fim, meu sincero agradecimento aos professores que gentilmente aceitaram participar da banca.

Resumo

Nas últimas décadas, a Inteligência Artificial (AI) por meio das mais recentes técnicas de Aprendizagem de Máquina tem impactado diversas áreas, especialmente a área da saúde, graças à sua crescente precisão e eficiência. Modelos preditivos de Aprendizagem de Máquina são promissores, mas é fundamental que sejam compreensíveis e explicáveis para os profissionais de saúde, aumentando sua aceitação e confiança. Com o avanço das ferramentas em nuvem, a oferta da IA como serviço (AIaaS) tem crescido, permitindo que organizações e ecossistemas se beneficiem do aprendizado de máquina para compor soluções. Contudo, a facilidade na criação e disponibilidade desses modelos de Aprendizagem de Máquina em nuvem traz à tona questões de transparência e interpretabilidade, principalmente em domínios sensíveis a esses aspectos como o domínio da saúde, que exige integridade e conformidade com princípios éticos e regulatórios. Esta pesquisa propõe uma arquitetura de referência baseada no paradigma AIaaS para construir modelos de aprendizado de máquina em nuvem, integrando explicabilidade para melhorar a tomada de decisões clínicas. A arquitetura HX-AIaaS visa propor uma arquitetura de referência para viabilizar explicabilidade em IA como serviço. Para viabilizar a arquitetura proposta foram implementados dois estudos de caso. O primeiro estudo de caso envolveu dados tabulares e foram usadas as técnicas de explicabilidade LIME, Anchor Rules e contrafactual. O segundo estudo de caso envolveu dados de imagens e envolveu as técnicas LIME e Grad-CAM. A arquitetura proposta revela-se promissora no contexto de criação de modelos de aprendizado de máquina e suas explicações no que diz respeito a transparência e interpretabilidade, o que favorece os processos de tomada de decisões clínicas.

Palavras-chave: Aprendizagem de Máquina, Explicabilidade, AIaaS, Inteligência Artificial

Abstract

In recent decades, Artificial Intelligence (AI) through the latest Machine Learning techniques has impacted various fields, especially healthcare, due to its growing precision and efficiency. Predictive Machine Learning models are promising, but it is crucial that they are understandable and explainable to healthcare professionals, increasing their acceptance and trust. With the advancement of cloud tools, the offering of AI as a service (AIaaS) has grown, allowing organizations and ecosystems to benefit from machine learning in developing solutions. However, the ease of creating and deploying these cloud-based Machine Learning models raises issues of transparency and interpretability, particularly in domains sensitive to these aspects, such as healthcare, which demands integrity and compliance with ethical and regulatory principles. This research proposes a reference architecture based on the AIaaS paradigm to build cloud-based machine learning models, integrating explainability to improve clinical decision-making. The H-XAIaaS architecture aims to propose a reference architecture to enable explainability in AI as a service. To make the proposed architecture feasible, two case studies were implemented. The first case study involved tabular data, and the explainability techniques LIME, Anchor Rules, and counterfactual were used. The second case study involved image data and employed the LIME and Grad-CAM techniques. The proposed architecture proves promising in the context of creating machine learning models and their explanations concerning transparency and interpretability, which supports clinical decision-making processes.

Keywords: Machine Learning, Explainability, AIaaS, Artificial Intelligence.

Lista de tabelas

Lista de ilustrações

Sumário

1 Introdução

Nas últimas décadas, temos testemunhado avanços significativos na área da Inteligência Artificial (IA), uma tecnologia que tem impactado diversas esferas do conhecimento, proporcionando melhorias e métodos de apoio à tomada de decisões nunca imaginadas.

Devido à sua precisão, eficiência, estabilidade e escalabilidade, os modelos preditivos estão sendo considerados como uma solução promissora para desafios na área da saúde (MERJULAH; CHANDRA, 2019). No entanto, é crucial que esses modelos sejam compreensíveis para serem aplicados efetivamente no contexto clínico. Isso se torna imprescindível para que os profissionais de saúde possam compreender o funcionamento e as bases das decisões tomadas pelos modelos.

As técnicas de explicabilidade desempenham um papel crucial nesse cenário, ao aumentar a aceitação e a confiança nos modelos de IA quando aplicados na prática clínica. Essa importância decorre da capacidade dessas técnicas em estabelecer conexões causais, proporcionando um entendimento mais aprofundado dos processos decisórios dos modelos (JUNG et al., 2023).

Conforme observado por (ASSADI et al., 2022), diversos outros obstáculos devem ser considerados para a implementação prática desses modelos. Entre esses desafios, destaca-se a necessidade de conjuntos de dados bem rotulados e de alta qualidade. Isso requer uma padronização nas técnicas de coleta de dados, que deve identificar e mitigar a presença de dados imprecisos ou incompletos. É crucial garantir que os conjuntos de dados de treinamento e teste sejam verdadeiramente representativos de diferentes populações.

Ademais, há outros fatores igualmente relevantes a serem ponderados, quando falamos sobre aplicações de inteligência artificial. De acordo com a natureza do seu problema, traz consigo desafios inerentes, quanto relacionados à manutenção e a integridade destes modelos.

Podemos citar como exemplo, a área da saúde. Que envolve a capacidade em fornecer explicações de maneira rápida e/ou em tempo real, que seja em conformidade com as necessidades clínicas. Além de respeitar os princípios éticos é uma consideração vital, incluindo a prevenção da discriminação de gênero e raça. A conformidade com as leis e regulamentos vigentes também é um aspecto incontornável, como é o caso da regulamentação da AI na União Europeia (EUROPEIA, 2024).

Diante das problemáticas identificadas e do avanço das ferramentas em nuvem, como a Google Cloud (Google)[1], Azure (Microsoft)[2] e AWS (Amazon)[3], a indústria visualizou a

[1] **Google Cloud Plataform**: <https://cloud.google.com/?hl=pt-BR>
[2] **Azure**: <https://azure.microsoft.com/pt-br/>
[3] **Amazon Web Services**: <https://aws.amazon.com/pt/free/>

robustez que tais gigantes da tecnologia possuem, devido à sua experiência e infraestrutura. A partir de tal fato, as grandes empresas estão se utilizando da sua robustez e oferecendo cada vez mais a Inteligência Artificial como serviço (do inglês, *Artificial Intelligence as a Service*, sigla AIaaS).

Além disso, uma infraestrutura na nuvem possibilita a integração de IA em aplicativos, permitindo que empresas sem tempo e recursos para desenvolver essas soluções internamente se beneficiem das capacidades avançadas de aprendizado de máquina e automação. Grandes empresas estão aproveitando suas infraestruturas para oferecer cada vez mais AIaaS, disponibilizando uma série de serviços que abrangem desde armazenamento até a construção de modelos. Esses modelos podem ser tanto pré-treinados quanto desenvolvidos do zero.

Tais modelos podem atender diferentes necessidades, que vão desde de tarefas de previsões, classificação, reconhecimento de objetos etc. O que atrai clientes de diferentes realidades, que podem consumir tais serviços sem dificuldades.

Esta pesquisa propõe uma arquitetura de referência baseada no padrão de AIaaS para a construção de modelos de aprendizado de máquina na nuvem, abrangendo desde a etapa de treinamento até a avaliação, e integrando uma arquitetura dedicada à explicabilidade. Essas abordagens têm potencial para aprimorar o diagnóstico e a tomada de decisões no âmbito da saúde, ao proporcionar maior transparência e confiança nos resultados gerados pelos sistemas de IA.

Dessa maneira, espera-se que a arquitetura proposta, denominada H-XAIaaS, promova melhorias significativas na aplicação de inteligência artificial em diversas indústrias, especialmente na saúde, através de uma plataforma que une robustez, explicabilidade e eficiência operacional.

1.1 Objetivo Geral

Objetivo deste trabalho é desenvolver uma arquitetura voltada para a área da saúde, integrando modelos de aprendizado de máquina em cenários clínicos reais, visando aprimorar a confiabilidade, a interpretabilidade e a eficácia desses modelos, além de oferecer um conjunto de práticas que garantam a transparência e a aplicação responsável dessas tecnologias no contexto médico.

1.2 Objetivos Específicos

Para alcançar o objetivo geral foram definidos objetivos específicos que podem ser identificados abaixo.

- **Objetivo 1**: Realizar uma análise da literatura acadêmica pré-existente com o

objetivo de identificar e avaliar arquiteturas, *frameworks* e protótipos cuja a finalidade sejam modelos de aprendizado de máquina aplicados ao mundo real na área da saúde;

- **Objetivo 2**: Identificar os componentes que farão parte da arquitetura de explicabilidade a ser proposta, fornecendo detalhes de sua execução;
- **Objetivo 3**:Desenvolver uma arquitetura de referência que integra práticas de H-XAIaaS voltada para o domínio da saúde, garantindo a transparência e a explicabilidade dos modelos empregados em diferentes cenários clínicos;
- **Objetivo 4**: Realizar dois estudos de caso utilizando conjuntos de dados de saúde reais, aplicando a arquitetura proposta para demonstrar como a adoção da arquitetura proposta, pode melhorar a confiabilidade dos modelos de *Machine Learning* e facilitar a interpretação por parte dos profissionais de saúde; ;
- **Objetivo 5**:Avaliar a eficácia da arquitetura em termos de explicabilidade e transparência, a partir dos módulos desenvolvidos para interpretabilidade dos modelos de *Machine Learning*, por meio de métricas específicas que mensuram a capacidade dos módulos em fornecer *insights* claros e compreensíveis sobre as decisões tomadas pelos modelos, visando assim atender aos requisitos fundamentais de interpretabilidade exigidos no contexto clínico da saúde.

1.3 Organização do Documento

O presente trabalho está dividido em seis capítulos distintos, cada uma abordando aspectos cruciais na integração de AIaaS e explicabilidade em contextos de saúde.

No Capítulo 2, são introduzidos os conceitos fundamentais e os arcabouços teóricos essenciais para a compreensão da pesquisa.

O Capítulo 3 apresenta uma revisão bibliográfica narrativa sobre os desafios das arquiteturas relacionadas ao aprendizado de máquina e à explicabilidade, além de discutir o papel dos usuários em relação a essas arquiteturas.

No Capítulo 4, é apresentada a arquitetura proposta, com uma descrição detalhada dos componentes que incorporam técnicas de explicabilidade, destacando os módulos que tornam os modelos de Machine Learning interpretáveis e transparentes.

O Capítulo 5 apresenta estudos de caso nos quais a arquitetura H-XAIaaS foi implementada, juntamente com os resultados obtidos.

No Capítulo 6, as conclusões da pesquisa são discutidas, assim como as perspectivas para trabalhos futuros.

2 Fundamentação Teórica

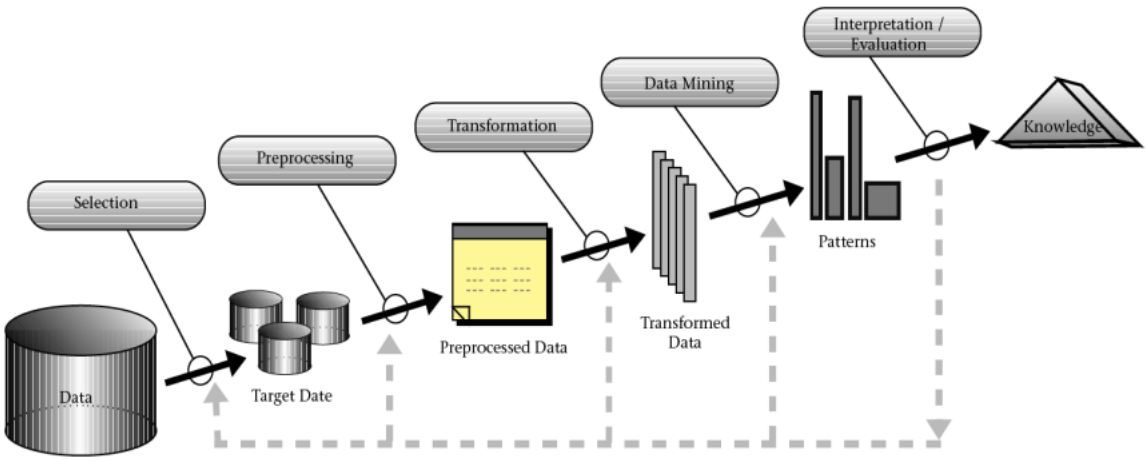
Este capítulo apresentará detalhes técnicos das atividades e resultados dos capítulos seguintes. Inicialmente, a seção 2.1 explora a descoberta de conhecimento a partir de bancos de dados e técnicas de mineração de dados. Em seguida, a seção 2.2 aborda os fundamentos da inteligência artificial. A seção 2.3 introduz conceitos essenciais relacionados ao aprendizado de máquina e redes neurais. Posteriormente, a seção 2.4 detalha aspectos da explicabilidade em modelos de aprendizado de máquina, complementada pela seção 2.5, que discute técnicas específicas de explicabilidade. A seção 2.6 é dedicada às métricas de explicabilidade. Além disso, a seção 2.7 destaca a necessidade e a importância da explicabilidade. Avançando para a seção 2.8, discute-se o conceito de conhecimento como serviço. Por fim, a seção 2.9 examina a Inteligência Artificial como Serviço (AIaaS).

2.1 Descoberta do Conhecimento em Banco de Dados

Descoberta do conhecimento em Banco de dados, (em inglês, *Knowledge Discovery in Databases*), pode ser descrito como processo de identificar padrões válidos, novos e potencialmente úteis e compreensíveis em dados, é o processo de extração não trivial, interativo de conhecimento implícito e com múltiplos estágios a partir de sistemas de banco de dados (FAYYAD, 2001), é um processo que envolve desde a preparação da base de dados até a apresentação do conhecimento deles extraído pelas técnicas de mineração.

A figura 1, exibe os cinco estágios para descoberta do conhecimento, propostos por

Figura 1 – : Estágios da Descoberta de Conhecimento em banco de dados. Fonte: (FAYYAD, 2001)



Fayyad (2001), nos quais são:

- **Seleção de Dados**: Nesta fase, os dados relevantes são selecionados a partir de várias fontes disponíveis. É importante garantir que os dados selecionados sejam representativos e adequados para análise.
- **Pré-processamento de Dados**: Nesta fase, os dados brutos são limpos e transformados para torná-los adequados para análise. Isso pode incluir a remoção de dados duplicados, a correção de erros de digitação e a normalização dos dados.
- **Transformação de Dados**: Nesta fase, os dados são transformados em um formato que pode ser usado para análise. Isso pode incluir a agregação de dados, a discretização de dados contínuos e a criação de novas variáveis.
- **Mineração de Dados**: Nesta fase, várias técnicas de mineração de dados são aplicadas aos dados transformados para extrair padrões e conhecimentos úteis. Isso pode incluir a identificação de associações, a classificação de dados e a previsão de valores.
- **Avaliação de Resultados**: Nesta fase, os resultados da mineração de dados são avaliados para determinar sua relevância e utilidade. Isso pode incluir a validação dos resultados e a identificação de possíveis limitações.
- **Apresentação de Resultados**: Nesta fase, os resultados da mineração de dados são apresentados de uma forma que possa ser facilmente compreendida e utilizada. Isso pode incluir a criação de relatórios, gráficos e visualizações.

2.1.1 Mineração de Dados

De acordo com Grossman, Hornick e Meyer (2002), o processo de dados para encontrar padrões, associações, mudanças, anomalias e estruturas estatísticas e eventos em um conjunto de dados é conhecido como mineração de dados (do inglês, *data mining*). Raramente o processo de descobertas de tais relações é de forma totalmente automatizada, sendo necessárias intervenções do analista .

Segundo Moxon (1996), de forma semelhante a Grossman, Hornick e Meyer (2002), define *data mining* como um conjunto de técnicas que visam explorar exaustivamente e revelar relações complexas entre grandes conjuntos de dados. De forma geral, o objetivo do *data mining* é explicar o conhecimento que está velado nas grandes massas de dados, permitindo assim a tomada de decisão.

Segundo QUONIAM (2001), podemos definir algumas técnicas em *data mining*:

- **Análise de Regras de Associação**, é um padrão de descoberta em dados transacionais que estabelece uma relação entre os itens, geralmente expressa como "se X,

então Y", capturando a associação de itens em um conjunto de dados, permitindo prever a ocorrência de um item com base na presença de outros.

- **Análise de Padrões Sequenciais**: Refere-se à identificação de padrões e relações sequenciais em conjuntos de dados onde a ordem das ocorrências é relevante. Isso envolve a descoberta de sequências frequentes, como eventos, ações ou transações que ocorrem em uma determinada ordem.
- **Classificação e Predição**: Processo de criar modelos que descrevem e categorizam dados em classes ou categorias predefinidas com base em características conhecidas, com o objetivo de prever a classe de objetos que ainda não foram classificados. É uma tarefa de aprendizado supervisionado, onde um algoritmo é treinado usando dados de treinamento que possuem rótulos ou classes atribuída
- **Análise de *Clusters***: Técnica que agrupa objetos ou dados semelhantes em grupos chamados *clusters* com base em características em comuns. Essa abordagem é comumente utilizada para examinar a estrutura interna dos dados e encontrar relações não óbvias entre eles. É uma abordagem não supervisionada, onde grupos não são pré-definidos, mas sim identificados pelo algoritmo com base nas suas similaridades.
- **Análise de *Outliers***: É a localização e análise de pontos de dados atípicos ou inconsistentes em um conjunto de dados. A análise e os resultados podem ser distorcidos por esses valores incomuns.

2.2 Inteligência Artificial

Segundo Russell e Norvig (2010) definem a inteligência artificial (IA) como, o estudo de agentes inteligentes que recebem percepções do ambiente e realizam ações. Cada a gente é implementado através de uma função que mapeia as percepções em ações, e abordamos diferentes maneiras de representar essas funções, seja por agentes, redes neurais e sistemas de teoria da decisão, buscando a construção de agentes inteligentes. De modo similar Marr (1977) define que a AI como identificação de problemas e suas soluções através do processamento de informações.

2.2.1 Aprendizado de Máquina

Aprendizado de Máquina (do inglês, *Machine Learning*), é um subcampo da AI que utiliza técnicas da estatística e matemática, para compreender padrões através de um conjunto de dados (BROWN, 2021). Mitchell (1997), define aprendizado de máquina como "Um programa de computador que aprende com experiência E em relação a alguma classe de tarefas T e a medida de desempenho P se seu desempenho em tarefas T, conforme medido por P, melhora com a experiência E".

Nos últimos anos, as técnicas de aprendizado de máquina têm recebido atenção significativa, impulsionadas pela proliferação de dados e pela crescente acessibilidade proporcionada pelo avanço do hardware. Esse cenário tem atuado como um verdadeiro catalisador para a área. Modelos de aprendizado de máquina estão ganhando cada vez mais destaque, especialmente na comunidade médica, onde surgem como soluções promissoras para diversos desafios. Isso é evidente pelo fato de que aproximadamente 86% das organizações de saúde já utilizam alguma forma de solução baseada em aprendizado de máquina (SIWICKI, 2017).

Algoritmos de aprendizado de máquina são treinados utilizando conjuntos de dados, e posteriormente tomam decisões com base em padrões e *insights* extraídos desses dados. Isso permite que esses sistemas se adaptem e aprimorem seu desempenho com base nas informações absorvidas (O'MAHONY et al., 2013).

O processo de aprendizagem de máquina é iniciado a partir de uma entrada de dados, chamada conjunto de treinamento, que representa a experiência, e são utilizados para aprender as relações entre as características dos dados e suas respectivas saídas. Após o treinamento do modelo com o conjunto de treinamento, a saída é geralmente outra aplicação capaz de realizar alguma tarefa específica. Esse programa é desenvolvido com base no conhecimento adquirido pelo modelo durante o treinamento.

Segundo Russell e Norvig (2010), a aprendizagem de máquina pode ser divida em três grupos:

- **Aprendizado Supervisionado**: A aprendizagem supervisionada é um tipo de aprendizado de máquina em que a máquina aprende sob supervisão. Na aprendizagem supervisionada, a máquina é treinada usando dados rotulados, o que significa que a resposta alvo já é conhecida. Os dados rotulados são fornecidos à máquina, que analisa e aprende a associação dos dados com base em suas características. Uma vez que a máquina tenha aprendido a partir dos dados rotulados, ela pode fazer previsões precisas quando novos dados são fornecidos a ela sem qualquer rótulo. A aprendizagem supervisionada pode ser dividida em dois tipos: classificação e regressão.

- **Aprendizado Não Supervisionado**: Técnica de aprendizado de máquina em que os modelos aprendem padrões e estruturas a partir de dados não rotulados, sem nenhum conhecimento ou orientação prévia. Ao contrário da aprendizagem supervisionada, a aprendizagem não supervisionada não depende de dados rotulados. Os algoritmos analisam os dados e identificam padrões, relacionamentos ou agrupamentos sem categorias ou classes predefinidas.

- **Aprendizado por Reforço**: Aprendizagem por reforço é um subconjunto do aprendizado de máquina em que um sistema é orientado por um agente, que aprende

por tentativa e erro. O objetivo do aprendizado por reforço é ensinar um agente a tomar decisões em um ambiente interativo, com base em recompensas e punições recebidas por suas ações.

Em aprendizado de máquina, um dos possíveis problemas é a generalização, na qual, em alguns casos o modelo tem um desempenho excelente no conjunto de treinamento, mas seu desempenho decai para dados nunca antes vistos. Tal problema é chamado de sobreajuste ou *overfitting*. Isso ocorre devido a várias razões, como o tamanho dos dados de treinamento sendo muito pequeno e não contendo amostras de dados suficientes para representar com precisão todas as possíveis entradas de dados.

Existem diversas formas de prevenir o problema de *overfitting*, como técnicas de validação cruzada, como a técnica de *k-folds*, na qual o modelo é treinado em subconjuntos de tamanho igual e avaliado k vezes. Cada vez é utilizado um subconjunto diferente como conjunto de validação e *folds* restantes como conjunto de treinamento. Após todas as iterações e a escolha de uma métrica de desempenho de cada *fold* são calculadas as médias para fornecer uma estimativa final do desempenho do modelo.

Técnicas de *data augmentation*, se referem ao aumento artificial do tamanho de conjunto de dados, onde aplica-se diversas transformações aos dados existentes. Por exemplo, em tarefas de classificação de imagens, é possível realizar inversões, rotações, redimensionamentos e deslocamentos nas imagens para criar novos exemplos de treinamento.

Em certos algoritmos é possível através de hiperparâmetros, utilizar reguladores, que de forma geral obriga os algoritmos de aprendizado de máquina a simplificar o modelo, adicionando através de um termo de penalização à função de perda, desencorajando o modelo de se ajustar muito aos dados de treinamento.

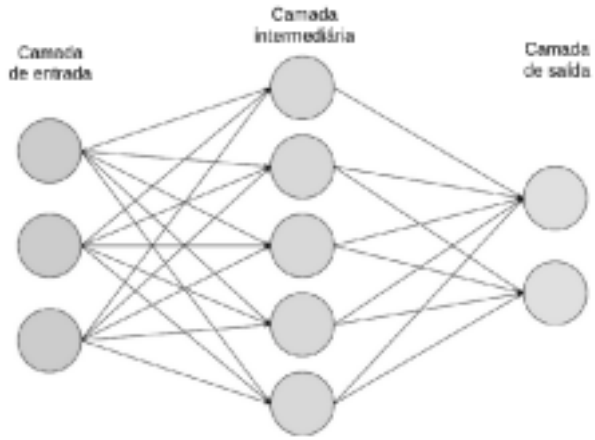
2.2.2 Redes Neurais Artificiais

Uma Rede Neural Artificial (RNA) é inspirada no funcionamento do cérebro humano. Sua estrutura consiste em nós interconectados, também conhecidos como neurônios, organizados em camadas interligadas (conexões), normalmente unidirecionais. Cada conexão entre neurônios possui um peso associado, que determina a força da conexão.

Esses pesos são ajustados durante o processo de treinamento visando otimizar o desempenho da rede. Cada neurônio aplica uma função de ativação à soma ponderada de suas entradas. Tal processo introduz a não-linearidade na rede, permitindo que a mesma aprenda padrões complexos e realize previsões não lineares. Redes Neurais Artificiais que possuem mais de uma ou mais camadas ocultas são denominadas de Redes Neurais Profundas (BRAGA; LUDERMIR; CARVALHO, 2000), conforme podemos visualizar na Figura 2.

Em Redes Neurais Artificiais com a arquitetura *feed-forward*, são unidirecionais

Figura 2 – Representação de uma rede neural



e acíclicas, que se utilizam da propagação direta para a calcular a saída da rede dado uma entrada. A entrada é passada pela camada da rede, e cada camada aplica uma transformação à entrada para produzir uma saída.

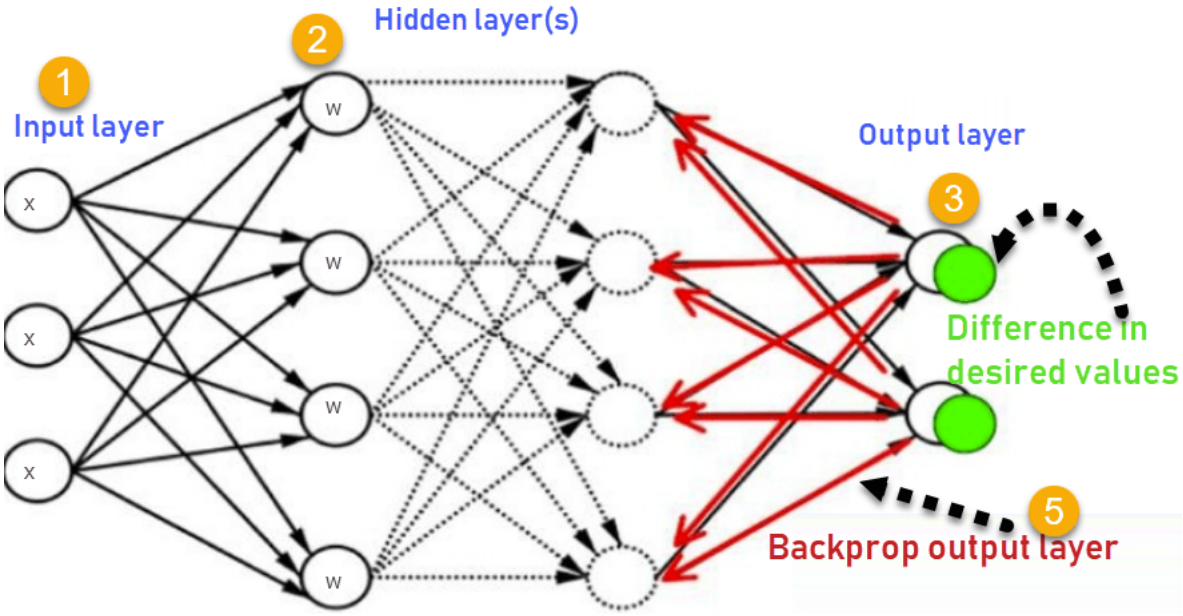
Uma outra técnica utilizada por RNAs é a técnica de *backpropagation*, método que visa atualizar os pesos da rede com base na taxa de erro obtida na época ou iteração anterior. É a essência do treinamento de redes neurais e é usado para ajustar os pesos de uma rede neural de forma a garantir taxas de erro mais baixas, tornando o modelo confiável ao aumentar sua generalização (BEEMAN, 2001), Figura 3.

Podemos definir os seguintes passos do algoritmo de *backpropagation*:

- Entradas X chegam através do caminho pré-conectado;
- As entradas X são modeladas usando pesos reais W. Os pesos são geralmente selecionados aleatoriamente;
- Calcula-se a saída para cada neurônio, da camada de entrada, passando pelas camadas ocultas até a camada de saída.
- Calcula-se o erro nas saídas. ErroB = Saída Real - Saída Desejada
- Retorna-se da camada de saída para a camada oculta para ajustar os pesos de forma a diminuir o erro.

A arquitetura da Rede Neural tem grande impacto no resultado do treinamento e com base no conjunto de dados fornecido sua performance pode ser reduzida. Na área da saúde um subtipo das RNA é bastante utilizado, e considerado o estado da arte para os problemas de imagens, as Redes Convolucionais. O crescente número de revisões da literatura comprovam tal fato (van der Velden et al., 2022; MIRANDA; ARYUNI; IRWANSYAH, 2016; SHEN; WU; SUK, 2017).

Figura 3 – Representação do algoritmo de *backpropagation* Fonte: (CLARKE, 2024).



As redes convolucionais se utilizam de modificações em suas estruturas internas a partir da implementação de operações de convolução nos dados de entrada. A convolução envolve o uso de *kernels* ou filtros para multiplicar entre o filtro e uma pequena região na entrada. Isso faz com que o filtro se desloque por toda a extensão dos dados. Cada filtro realiza uma operação de convolução, calculando o produto escalar entre seus pesos e uma pequena região dos dados de entrada. Esse processo ajuda a extrair características locais e capturar relações espaciais. A rede pode aprender caracteristicas locais e invariantes, que podem ser relevantes independentemente de sua localizacao na entrada (O'SHEA; NASH, 2015).

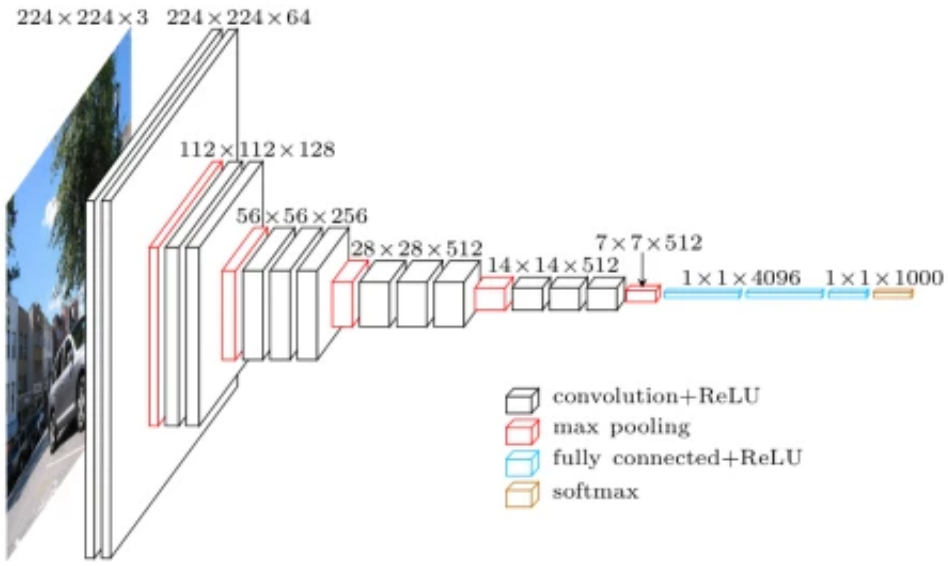
Após a operação de convolução, uma função de ativação é aplicada, elemento a elemento, para introduzir não-linearidade, por exemplo a função ReLU (*Rectified Linear United*). Após a função de ativação, a camada de *pooling* para reduzir as dimensões das *feature maps*. O *pooling* reduz as dimensões espaciais, enquanto retém as informações mais importantes.

De forma geral, as redes neurais, em conjunto com algoritmos de aprendizado de máquina, permitem encontrar soluções complexas que são difíceis para métodos tradicionais Huang, Zhu e Siew (2006). Devido sua capacidade de alta generalização, e acurácia, segue como um dos principais métodos para suporte à decisão.

2.2.3 *Visual Geometry Group* - VGG19

VGG19, é uma variante do VGG Model, Figura 4, é um acrônimo para *Visual Geometry Group* (Grupo de Geometria Visual) desenvolvida pela Universidade Oxford (SIMONYAN; ZISSERMAN, 2015), é um modelo de rede convolucional (CNN), na qual,

Figura 4 – Arquitetura VGGG19. Fonte: (SIMONYAN; ZISSERMAN, 2015),.



contém 19 camadas, sendo 16 camadas de convolução, 3 camadas totalmente conectadas, 5 camadas de Max *Pooling* e uma camada de *output* da predição, a figura 4 exibe a arquitetura VGG, com a função de ativação '*softmax*', função utilizada especialmente para problemas de classificação, na qual a partir de uma entrada retorna um conjunto de probabilidades sobre as diferentes classes de saída.

2.2.4 Métodos Baseados em Árvores

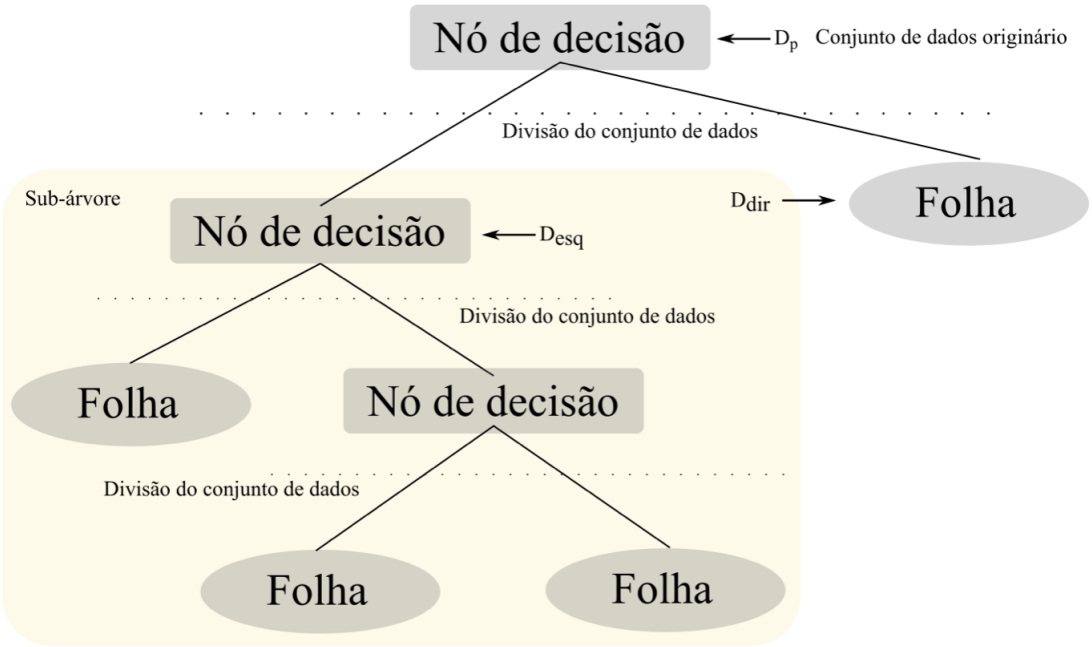
Entre os diversos algoritmos de aprendizado de máquina, um dos mais populares é o de árvore de decisão. Utilizado tanto para tarefas de regressão quanto de classificação, tal algoritmo se utiliza de uma estratégia gulosa para tomada de decisões. Significa que a cada etapa o algoritmo toma a decisão ótima aproximada com base nos dados disponíveis e representa esta decisão em forma de árvore.

Árvores de decisão classificam instâncias ao organizá-las descendo pela árvore desde a raiz até algum nó folha, que fornece a classificação da instância. Cada nó na árvore especifica um teste de algum atributo da instância, e cada ramo descendente desse nó corresponde a um dos valores possíveis para esse atributo.

A decisão de dividir em cada nó é tomada de acordo com uma métrica denominada de entropia, que caracteriza a impureza de um conjunto de exemplos, de modo geral a entropia mede quantos *bits* de informação são necessários para codificar a classificação de um dado, ou seja, a entropia mede o quão nossos dados estão desorganizados (MITCHELL, 1997), figura 5.

Onde quanto maior a entropia menor é o ganho de informação e vice-versa. O algoritmo funciona dividindo inicialmente o conjunto de treinamento em dois subconjuntos. Uma vez que o algoritmo tenha dividido com sucesso o conjunto de treinamento, ele divide

Figura 5 – Exemplo de funcionamento da árvore de decisão.



os subconjuntos usando a mesma lógica e assim por diante, de forma recursiva.

O processo de divisão se encerra quando atinge a profundidade máxima ou não consegue encontrar uma divisão que reduza a impureza. Por se tratar de um modelo de funcionamento bastante simples, o mesmo é bastante interpretável por humanos (RUSSELL; NORVIG, 2010).

Um algoritmo bastante utilizado é de Florestas Aleatórias (do inglês, *Random Forest*) (BREIMAN, 2001). O mesmo é denominado de método *ensemble*, que produz como modelo resultante várias árvores de decisões combinadas. O classificador Floresta Aleatória é um algoritmo de aprendizado de máquina que utiliza várias árvores de decisão para fazer previsões, onde cada árvore de decisão é treinada em um subconjunto aleatório dos dados de treinamento, e as previsões das árvores são então combinadas para fazer a previsão final.

Uma outra forma de combinar árvores de decisão é conhecido pelo algoritmo de *ExtraTrees* ou *Extremely Randomized Trees* (GEURTS; ERNST; WEHENKEL, 2006). A diferença entre o algoritmo *Random Forest*, é que o algoritmo do *Random Forest* utiliza a técnica de *bootstraping* (amostragem com reposição) para a criação de subconjuntos de dados de treinamentos em cada árvore, enquanto o algoritmo utiliza todo o conjunto de dados e adiciona aleatoriedade no processo de seleção de atributos.

Onde o algoritmo *ExtraTrees*, adiciona além do limiar na divisão de cada nó, um fator de aleatoriedade para dividir as variáveis, ao invés de utilizar de um algoritmo guloso para encontrar o ponto de divisão ótimo.

2.2.5 XGBoost - eXtreme Gradient Boosting

Outro algoritmo que ganhou bastante popularidade nos últimos anos é o XGBoost (*eXtreme Gradient Boosting*) (CHEN; GUESTRIN, 2016), uma generalização do algoritmo *Gradient Boost Machines*.

O XGBoost busca minimizar a função de custo por meio das regularizações Lasso (L1) e Ridge (L2), ajudando a prevenir o *overfitting* e melhorar a generalização do modelo, o que é essencial ao lidar com grandes conjuntos de dados e alta dimensionalidade. Este método utiliza aprendizado em conjunto, combinando previsões de vários modelos "fracos" (do inglês, *weak learners*).

Um modelo é denominado de fraco em aprendizado de máquina, quando desempenha uma tarefa melhor do que adivinhação aleatória em uma determinada tarefa, atingindo uma precisão um pouco acima de 50%, são geralmente caracterizados por sua simplicidade e baixa complexidade, por exemplo, árvores de decisão rasas (FREUND; SCHAPIRE, 1995).

Um fator de destaque do XGBoost é a validação cruzada nativa, que permite encerrar o treinamento quando não há mais benefícios significativos. Além disso, o XGBoost se beneficia do paralelismo, utilizando múltiplos núcleos de processamento

2.2.6 AutoML - *Automated Machine Learning*

O aprendizado de máquina automatizado abrange um conjunto de técnicas e métodos que simplificam o desenvolvimento de modelos de machine learning, tornando esse processo acessível a usuários sem expertise técnica. Isso é possível por meio da automação de diversas etapas do fluxo de criação de modelos, como o tratamento de dados, a seleção de hiperparâmetros e o manejo de *outliers* etc (BARATCHI M., 2024).

Um dos grandes desafios ao trabalhar com modelos de aprendizado de máquina é que várias etapas demandam intervenção humana, o que pode influenciar diretamente o desempenho do modelo. Dessa forma, ao aplicar técnicas de AutoML, essas etapas são abstraídas, facilitando o processo de construção do modelo.

A abordagem visa, a geração do modelo com base no conjunto de dados fornecido pelo usuário, assim obtendo as melhores métricas de acordo com a tarefa desejada (classificação ou regressão). Além de garantir uma boa generalização através de dados não vistos no conjunto de treinamento.

Um dos fundamentos de AutoML é a Busca espacial e a busca algorítmica que estão correlacionados a todas as possíveis escolhas de design, que o modelo pode utilizar. Quando analisamos hiperparâmetros, a depender do modelo, temos parâmetros categóricos e numéricos, cada um com seu respectivo valores que podem ser alocados.

Logo, o espaço de busca se torna mais complexo, visto que o precisa realizar uma

busca exaustiva com a finalidade de encontrar os melhores parâmetros em relação a performance do modelo, embora exista estratégias mais sofisticadas do que uma busca aleatória ou em grade (BARATCHI M., 2024).

2.2.7 Métricas de Avaliação de modelos

Após um modelo treinado, é necessário avaliar a performance do modelo e entender o seu desempenho, para tal, temos diferentes métricas de avaliação de desempenho. Ao usar diferentes métricas para a avaliação de desempenho, podemos mitigar possíveis vieses e erros com dados não vistos.

Por exemplo, sem fazer uma avaliação adequada do modelo de aprendizado de máquina usando diferentes métricas e dependendo apenas da precisão, isso pode levar a um problema quando o respectivo modelo é implantado em dados não vistos e resultar em previsões ruins.

Logo, é de fundamental importância utilizar diferentes métricas de avaliação, pois oferece uma visão mais abrangente do modelo, além de capturar diferentes aspectos do seu comportamento. De acordo com Castro e Ferrari (2017), em problemas de classificação binária, predições podem ter quatro possíveis classes são:

- **Verdadeiro positivo (VP)**: quando o método diz que a classe é positiva e, ao verificar a resposta, vê-se que a classe era realmente positiva;
- **Verdadeiro negativo (VN)**: quando o método diz que a classe é negativa e, ao verificar a resposta, vê-se que a classe era realmente negativa;
- **Falso positivo (FP)**: quando o método diz que a classe é positiva, mas ao verificar a resposta, vê-se que a classe era negativa;
- **Falso negativo (FN)**: quando o método diz que a classe é negativa, mas ao verificar a resposta, vê-se que a classe era positiva;

Com base nessas quatro categorias, são derivadas as principais métricas utilizadas para a avaliação de modelos:

- **Precision (Precisão)**: Proporção de exemplos classificados corretamente com positivos em relação a todos os exemplos classificados como positivos (verdadeiros positivos + falsos positivos)

$$\text{Precisão} = \frac{VP}{VP + FP}$$

- **Recall (Revogação/sensibilidade)**: Proporção de exemplos classificados corretamente como positivos, em relação a todos que são de fato positivos (verdadeiros

positivos + falsos negativos)

$$\text{Revogação} = \frac{VP}{VP + FN}$$

- **Acurracy (Acurácia)**: Proporção de exemplos classificados corretamente em relação a todos os exemplos.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}$$

- **F1-Score** : Média harmônica entre a precisão e a revogação. É útil ao lidar com conjuntos de dados desbalanceados, pois a precisão pode ser uma métrica enganosa.

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Revogação}}{\text{Precisão} + \text{Revogação}}$$

Uma outra medida bastante utilizada é a Curva de Característica de Operação do Receptor (ROC). A análise ROC (*Receiver Operating Characteristic*), é um técnica utilizada para visualizar e selecionar classificadores com base em seu desempenho. Historicamente o gráfico ROC tem sido utilizado para representar o comparativo entre taxa de acerto e taxas de falsos positivos entre os classificadores (SWETS; DAWES; MONAHAN, 2000).

A curva ROC é um gráfico que plota os verdadeiros positivos no eixo Y em função da taxa de falsos positivos, eixo X. (SPACKMAN, 1989) foi o primeiro a adotar os gráficos ROC, na área de aprendizado de máquina, que demonstrou o seu valor na comparação entre diferentes métodos. O modelo ideal terá uma curva ROC que se aproxima do canto superior esquerdo do gráfico, indicando alta taxa de verdadeiros positivos e uma baixa taxa de falsos positivos.

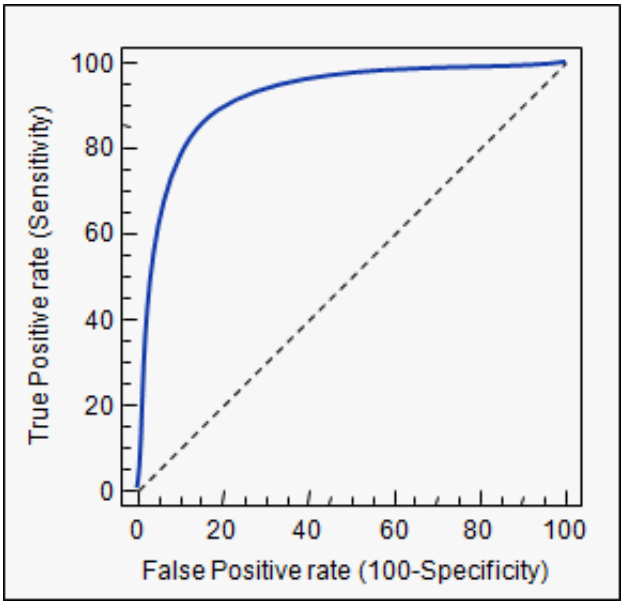
Existe uma métrica derivada da curva ROC, Área Sob a Curva ROC (AUC), que mede a capacidade do modelo em distinguir as classes, variando entre 0 a 1. Sendo, quanto maior o AUC maior a sua capacidade de distinguir entre as classes, como podemos visualizar na figura 6.

2.3 Explicabilidade

Sistemas de Aprendizado de Máquina, se utilizam de padrões baseados nos dados de entrada. Os sistemas de IA extraem padrões dos dados de entrada e obtêm percepções com base no conhecimento treinado. Devido à sua eficácia, as técnicas de aprendizado de máquina estão sendo usadas com mais frequência no campo médico (MERJULAH; CHANDRA, 2019).

Contudo, alguns cuidados precisam ser tomados. Segundo Caruana et al. (2015), em seu estudo onde desenvolve uma rede neural artificial para prever quais pacientes com pneumonia deveriam ser internados, e quais poderiam ser tratados em regime ambulatorial,

Figura 6 – Representação da Área sob a Curva



os estudos inicialmente foram promissores, possuindo métricas melhores que as abordagens tradicionais.

Entretanto, o modelo inferiu que pacientes com asma apresentavam um risco reduzido à morte por pneumonia do ponto de vista médico, isso não é lógico, pois, na prática médica, pacientes asmáticos geralmente têm um risco elevado de complicações, ocasionando a desistência da adoção de tal sistema devido que seu uso poderia ser danoso clinicamente.

Tal exemplo de saída gerada pelo estudo de Caruana et al. (2015) ilustra o problema de interpretação, e a necessidade de que modelos não sejam apenas precisos mas também explicáveis, para que os profissionais da saúde possam confiar nas previsões e tomar as melhores decisões de tratamento aos pacientes.

Devido a sua notável precisão modelos cada vez mais complexos, estão sendo cada vez mais utilizados, devido a complexidade do algoritmo não é possível entender o funcionamento interno do modelo e de como a decisão é tomada, denominados assim modelos "caixa-preta".

A capacidade de fornecer explicações sobre modelos "caixas pretas", pode tornar o modelo mais transparente e confiável, promovendo uma maior confiança ao consumidor das predições, além de uma maior evidência quando houver erros.

Diante de tal necessidade, surge a área de XAI (sigla para *Explainable Artificial Intelligence*, ou Inteligência Artificial Explicável), campo multidisciplinar que abrange as áreas de ciência da computação, engenharia da computação a psicologia, a partir de conceitos algorítmicos, psicológicos e ciência cognitiva. XAI visa fornecer decisões de IA explicáveis, permitindo que usuários sem experiência em aprendizado de máquina compreendam o comportamento do modelo.

2.3.1 O que é interpretabilidade?

A interpretabilidade é definida como, a medida em que os humanos podem entender como um modelo de aprendizado de máquina funciona. Um modelo com alto nível de interpretabilidade pode ser explicado de uma forma que qualquer um possa entender, independentemente de suas habilidades técnicas. Isso é importante porque permite que as pessoas confiem nos modelos de aprendizado de máquina e os usem com confiança (MASÍS, 2021).

De modo similar Doshi-Velez e Kim (2017) definem interpretabilidade no contexto de aprendizado de máquina como a habilidade de explicar ou tornar compreensível para os seres humanos, oferecendo-lhes um senso do mecanismo por trás do funcionamento do modelo.

Muitos modelos de aprendizado de máquina são inerentemente mais difíceis de entender simplesmente devido a suas propriedades matemáticas e arquiteturais inerentes, além de tal fato, outros processos podem aumentar a complexidade dos modelos e torná-los menos interpretáveis, como quais dados foram utilizados no modelo, processos de *feature engineering*, escolha de hiperparâmetros etc.

A forma de interpretabilidade compreensível pela linguagem humana, são o conjunto de técnicas que permitem que o usuário compreenda o processo decisório sem a necessidade de conhecimento especializado em matemática ou em computação, como explicações em linguagem natural ou em forma de gráficos, por exemplo.

Já a interpretabilidade baseada em modelos matemáticos, é o conjunto de técnicas que utilizam de modelos matemáticos para revelar os mecanismos internos de um modelo de aprendizado de máquina, por exemplo, a análise de pesos ou coeficientes de um modelo para entender sua importância a partir dos dados de entrada, ou a visualização de representações internas de um modelo para entender como o mesmo processa as informações.

2.3.2 O que é explicabilidade?

No artigo proposto por Langer et al. (2021), denominado de "*What Do We Want From Explainable Artificial Intelligence (XAI)? A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research*", definem explicabilidade como uma área multidisciplinar de pesquisa que se concentra no desenvolvimento de métodos de explicação e criação de sistemas artificiais compreensíveis para usuários humanos, tornando transparente o processo de tomada de decisão do modelo.

De forma semelhante, Gunning D.; Vorm e Turek (2021) conceitua XAI como o processo de tornar o comportamento de sistemas de inteligência artificial mais acessível aos seres humanos, proporcionando explicações. Os sistemas de XAI devem demonstrar a capacidade de explicar suas habilidades e compreensão, bem como revelar as informações

pertinentes nas quais baseiam suas ações.

(BIRAN; COTTON, 2017) definem a distinção entre interpretabilidade e explicabilidade, respectivamente, como: o quanto um observador consegue entender uma decisão e o método como um observador pode obter essa compreensão. No entanto os autores Tjoa e Guan (2020) tratam os termos como inigualáveis, trazendo a área da interpretabilidade como uma tentativa de explicar decisões de algoritmos, descobrir padrões dos mecanismos internos de algoritmos. Nesse sentido eles dividem as formas de interpretabilidade entre aquelas compreensíveis facilmente pela linguagem humana e por meio de modelos matemáticos.

Em resumo, a diferença entre explicabilidade e interpretabilidade é que a explicabilidade busca ir mais a fundo no quesito de transparência. Ela exige explicações mais amigáveis para o funcionamento interno de um modelo e seu processo de treinamento, e não apenas na inferência, podendo se estender ao projeto do algoritmo e à transparência algorítmica. Para garantir uma boa explicabilidade, é necessário possuir entendimento de todo o processo decisório, que pode ser dividido nos seguintes aspectos:

- **Transparência do modelo:** Capacidade de explicar passo a passo como um modelo é treinado.
- **Transparência do Projeto**: Capacidade de explicar as escolhas realizadas, como a arquitetura do modelo e seus hiperparâmetros.
- **Transparência Algorítmica**: Capacidade de entender como um modelo toma as decisões, onde o processo de tomada de decisão seja facilmente compreendido com base em qualquer saída a partir da entrada (FIELDING R.T; RICHARD, 2000a).

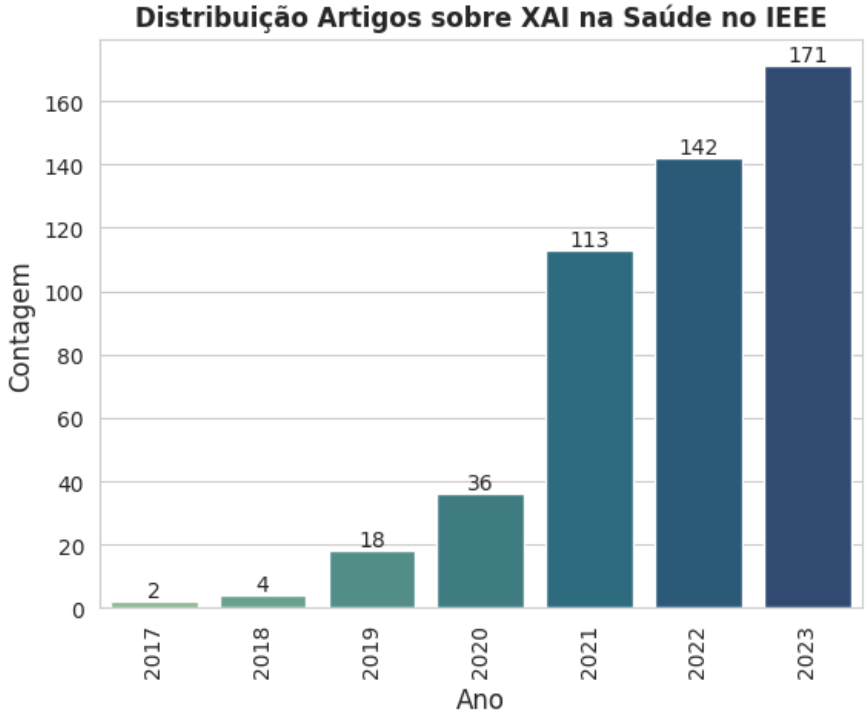
2.3.3 Por que modelos explicativos na saúde?

Nos últimos anos, a inteligência artificial (IA) tem influenciado significativamente diversas áreas do conhecimento, inclusive a área da saúde. Esse impacto está associado ao crescimento exponencial da disponibilidade de dados e ao aumento do poder computacional. Como resultado, a IA tem o potencial de melhorar a qualidade de vida e o bem-estar das pessoas, especialmente por meio do aprimoramento dos diagnósticos clínicos.

Ao utilizar modelos de aprendizado de máquina para resolver problemas na área da saúde, esses sistemas têm se mostrado equivalentes ou até superiores a especialistas humanos. Contudo, em ambientes clínicos reais, apresentam altas taxas de falsos positivos.

Essa limitação está relacionada ao uso de modelos denominados de "caixa-preta", que embora ofereçam benefícios relacionados a sua performance (MERJULAH; CHANDRA, 2019), dificultam a explicação do processo de tomada de decisão. Além dos desafios técnicos e éticos, a falta de compreensão sobre o diagnóstico gerado pela IA impede a garantia

Figura 7 – Distribuição de Artigos de XAI na saúde no portal IEEE



de que o diagnóstico reflete com precisão a realidade (BHARATI; MONDAL; PODDER, 2024).

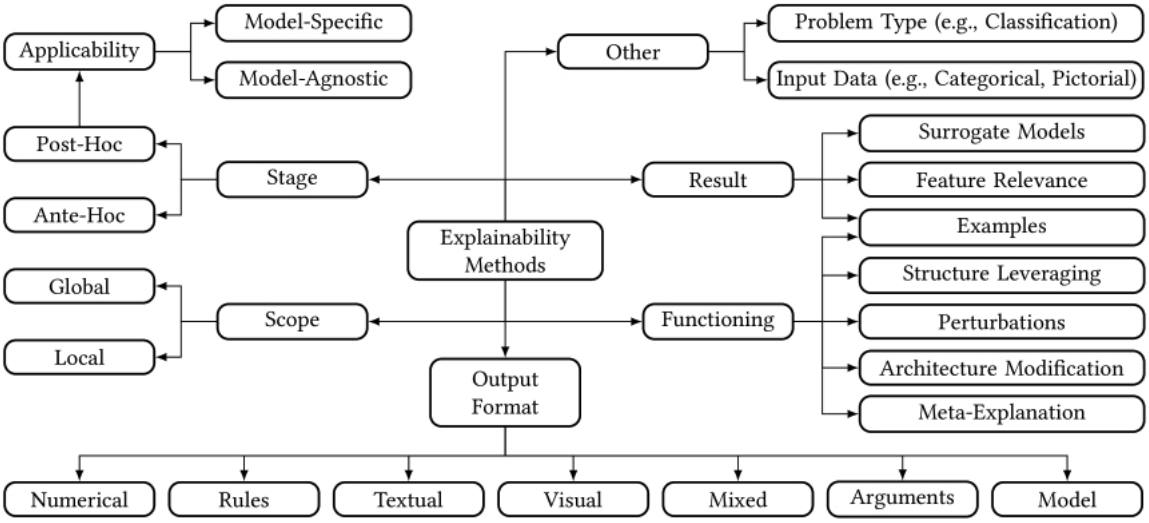
Logo na área da saúde, as técnicas de XAI são essenciais porque podem aumentar a confiança nos sistemas de IA e melhorar sua adoção na prática clínica. A partir de um melhor entendimento do processo decisório do modelo, as técnicas de XAI podem gerar hipóteses sobre a causalidade com o objetivo de aumentar a aceitação e a confiança Jung et al. (2023). Outro fator é reduzir e identificar vieses nos dados, pois os dados médicos podem conter vieses ocultos que levam a erros sistemáticos.

2.3.4 Taxonomia da Explicabilidade

A área de pesquisa em XAI está em constante crescimento, como evidenciado pelo aumento significativo no número de artigos publicados, conforme demonstrado na Figura 7 Devido à natureza relativamente recente e em constante evolução desse campo, uma ampla gama de métodos e abordagens estão sendo propostos e desenvolvidos. Isso apresenta um desafio para pesquisadores e profissionais, pois torna complexa a compreensão completa desse campo de estudo.

Com o intuito de proporcionar uma perspectiva estruturada dos métodos e conceitos relacionados à XAI (Explicabilidade em Inteligência Artificial), vários autores têm apresentado diversas abordagens para a taxonomia em XAI (ANGELOV et al., 2021; BELLE; PAPANTONIS, 2021; LANGER et al., 2021; MCDERMID et al., 2021). No entanto, cada um desses trabalhos traz interpretações distintas do conceito de explicabilidade, o que

Figura 8 – Taxonomia da Explicabilidade. Fonte: (SPEITH, 2022)



pode criar desafios para os pesquisadores. Por outro lado, a proposta de Speith (2022) oferece uma visão ampla que auxilia na redução das divergências entre os diferentes autores, contribuindo para uma compreensão mais coesa do campo e apresentando características relevantes para a modelagem por meio da XAI, como vista na Figura 8.

No contexto da Inteligência Artificial Explicável (XAI), a aplicabilidade refere-se à variedade de métodos que podem ser utilizados com diferentes tipos de modelos de aprendizado de máquina. Essas abordagens podem ser divididas em duas categorias principais: agnósticos ao modelo (*model-agnostic*), que são técnicas aplicáveis a qualquer tipo de modelo de aprendizado de máquina, e específico ao modelo (*model-specific*), que são técnicas adaptadas a um tipo particular de modelo (SPEITH, 2022; MOLNAR, 2022).

Além disso, ao considerar o estágio na taxonomia proposta por Speith (2022), encontramos duas abordagens distintas: *ante-hoc* e *post-hoc*. Os métodos *ante-hoc* são aplicados durante o processo de design e treinamento do modelo de aprendizado de máquina, enquanto os métodos *post-hoc* são aplicados após o modelo ter sido treinado.

No que diz respeito à abordagem de escopo, os métodos de XAI podem ser divididos em dois subgrupos: globais e locais. Os métodos locais têm como foco a explicação do comportamento do modelo em relação a previsões individuais, identificando os fatores que mais contribuíram para cada previsão. Por outro lado, os métodos globais têm como objetivo explicar o comportamento do modelo como um todo (LIPTON, 2017; MOLNAR, 2022).

Com base no método de explicabilidade escolhido, é possível apresentar as explicações em diferentes formatos, como formatos visuais, textuais, numéricos, entre outros. A seleção do formato de saída adequado pode ter um impacto significativo na utilidade e na compreensão das explicações, uma vez que os usuários podem ter diferentes necessidades e

preferências em relação à forma como desejam consumir essa informação.

Por exemplo, de acordo com um estudo realizado por Vilone e Longo (2020), afirmam que saídas numéricas são mais adequadas para usuários especializados do que para leigos. Além disso, um estudo conduzido por Diprose et al. (2020), indica que aproximadamente 66% dos especialistas preferem uma explicação local. Isso se deve ao fato de que tais explicações são mais intuitivas, fáceis de compreender e podem estar alinhadas com os processos de tomada de decisão de médicos, que geralmente envolvem a formulação de hipóteses sobre o diagnóstico do paciente.

2.3.5 Abordagens de explicabilidade

Visando aumentar a explicabilidade e transparência para diferentes tipos de modelos, estudos têm sido realizados que envolvem diversos tipos de abordagens, seja através de extração de regras que possuem o objetivo de ser interpretáveis por seres humanos a partir de modelos complexos, ou a partir da visualização de representações internas e importância de propriedades dos algoritmos. Bhattacharya et al. (2019) afirma que existem quatro abordagens diferentes de métodos de explicabilidade, que diferem de acordo com o tipo de modelo e dados.

2.3.5.1 Extração do Conhecimento

Visa fornecer explicabilidade através da extração de informações significativas a partir de dados ou do próprio modelo, auxiliando na identificação de padrões presentes, buscando entender a completude dos dados. Tal método se utiliza de abordagens estatísticas, seja por sumários, testes de hipóteses, estatísticas descritivas, análises univariadas, multivariadas etc.

2.3.5.2 Visualização de Resultados

A abordagem pós-processo de treinamento do modelo visa, na maioria das vezes, comparar várias possibilidades de resultados. Isso pode ocorrer com a utilização de técnicas de análise de componentes principais (PCA) para problemas de classificação, modelos de regressão ou séries temporais para visualizar intervalos de confiança do modelo. No entanto, como abordado por (DAS et al., 2020), modelos extremamente complexos, sejam da própria arquitetura do modelo ou do conjunto de seus componentes.

2.3.5.3 Baseados em Influência

Abordagem mais utilizada por diferentes técnicas de explicabilidade. Tal abordagem visa entender como certas características presentes nos dados impactam no processo da tomada de decisão do modelo, seja este impacto positivo ou negativo, uma abordagem comum é a técnica de Importância de atributos (do inglês, *Feature Importance*).

2.3.5.4 Baseados em Exemplo

O método mais acessível para usuários não técnicos, busca um contexto ou exemplo para ilustrar a explicação para o usuário. Isso permite que o usuário realize a correlação da saída do modelo com o mundo real.

Os métodos baseados em exemplos funcionam a partir de instâncias particulares de um conjunto de dados, geralmente métodos baseados em exemplos funcionam bem quando características carregam mais contexto, como imagens e texto. Para dados tabulares existe um desafio, devido que dados tabulares podem conter centenas de colunas, o que diluem tais características (MOLNAR, 2022).

2.3.6 Métodos de XAI

Os métodos de explicabilidade diferenciam-se em vários aspectos. Primeiramente, algumas abordagens são projetadas para funcionar especificamente com determinados tipos de dados, como técnicas que lidam exclusivamente com dados tabulares e/ou outras voltadas para o domínio de imagens.

Outro aspecto relevante é a forma como a explicabilidade é apresentada ao usuário. Algumas técnicas se utilizam gráficos ou mapas de calor, que destacam visualmente os atributos/regiões de uma imagem com maior importância, constituindo ferramentas de explicação visual. Outras abordagens fornecem explicações por meio de regras.

Neste trabalho, focaremos nos seguintes métodos de explicabilidade: Local Interpretable Model-Agnostic Explanation (LIME), Anchor Rules e explicações contrafactuais. A escolha dessas abordagens baseia-se em dois critérios principais: primeiro, são amplamente utilizadas na academia, sendo consideradas estado da arte para a área de explicabilidade; segundo, oferecem diferentes perspectivas para compreender a explicabilidade, enriquecendo a análise e interpretação dos modelos de aprendizado de máquina.

2.3.6.1 LIME (*Local Interpretable Model-Agnostic Explanation*)

A técnica LIME, abreviação de Local Interpretable Model Agnostic Explanation, método proposto por Ribeiro, Singh e Guestrin (2016). A ideia do algoritmo é fornece explicações locais e interpretáveis para previsões individuais feitas por qualquer modelo de aprendizado de máquina, através de uma perturbação/permutação entre as variáveis para gerar a explicação

A ideia do LIME é bastante intuitiva, na qual, o LIME testa a previsão do modelo a partir da perturbação de um conjunto de dados de entrada, a partir deste novo conjunto de amostras perturbadas o LIME treina um modelo interpretável, geralmente uma regressão lasso, que pondera a proximidade das instâncias perturbadas com o dado real.

Por ser um método local, a técnica não provê boas explicações globais, esse método local, é denominado de fidelidade local, dada pela seguinte fórmula (MOLNAR, 2022).

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

A função de geração da explicação é definida da seguinte forma: dada uma instância x, o modelo interpretável g minimiza uma função de perda L (do inglês, *loss function*), como por exemplo o erro quadrático médio, que mede quão próxima a explicação está da previsão original do modelo, como uma rede neural. Ao mesmo tempo, a complexidade do modelo $\Omega(g)$ é mantida baixa, o que significa a inclusão de menos variáveis. O conjunto G define todas as explicações possíveis. A medida de proximidade π_x determina o tamanho da vizinhança ao redor da instância x que consideramos para a explicação, e maneira geral, o algoritmo funciona da seguinte forma:

1. Dada uma instância (amostra) de dados;
2. Realiza-se a perturbação desses dados, gerando um novo conjunto a ser utilizado para a predição;
3. Atribuem-se pesos para os resultados, de acordo com a proximidade das amostras em relação à base de referência (sem a perturbação);
4. Treina-se um modelo usando o conjunto de dados perturbados no item 2;
5. Explicam-se as predições utilizando um modelo interpretável gerado no item 4.

A técnica LIME é utilizada tanto para o contexto de dados tabulares quanto para imagens. Em ambos os casos, a explicação de saída é apresentada de forma visual, permitindo que os usuários compreendam de maneira intuitiva quais atributos presentes nos dados influenciam as previsões do modelo. No caso de dados tabulares, a explicação é tipicamente exibida em gráficos de barras, onde cada barra representa a contribuição de uma variável específica para a predição. Para imagens, a técnica utiliza mapas de calor que destacam as áreas da imagem que mais impactam a decisão do modelo, facilitando a interpretação dos resultados.

2.3.6.2 Anchor Rules

Método de âncoras busca gerar explicações individuais para qualquer modelo denominado "caixa-preta". Uma regra se define como ancora, onde uma mudança entre os valores das variáveis não afetam a previsão. Tal método se utiliza de aprendizado por reforço que realiza a busca em um grafo, o método foi proposto por (RIBEIRO; SINGH; GUESTRIN, 2018).

O método *anchor rules* se utiliza de perturbações para gerar explicações locais, contudo ao invés de utilizar um modelo interpretável, as explicações são geradas como regras "se-então", uma das grandes vantagens deste método é que as regras são delimitadas, as âncoras incluem noção de cobertura, declarando precisamente a quais outras instâncias, possivelmente não vistas, elas se aplicam.

O método âncora é formalmente definido pela seguinte expressão.

$$E_{D_x(z|A)}\left[1_{\hat{f}(x)=\hat{f}(z)}\right] \geq \tau, \mathcal{A}(x) = 1$$

Dada uma instância x a ser explicada, é necessário identificar uma regra ou âncora A que se aplique a x. Essa regra deve garantir que a mesma classe prevista para x seja também prevista para pelo menos uma fração τ dos vizinhos de x onde A seja aplicável. A precisão da regra é determinada pela avaliação desses vizinhos, ou seja, das perturbações (seguindo a distribuição $D_x(z \mid A)$), utilizando o modelo de aprendizado de máquina fornecido, sendo essa precisão indicada pela função

$$1_{\hat{f}(x)=\hat{f}(z)}(MOLNAR, 2022).$$

A técnica de *Anchor Rules* é aplicada em contextos de dados tabulares e fornece uma saída textual que exibe tanto a precisão da regra quanto a cobertura, que representa o percentual da base de dados para o qual a regra é aplicável.

2.3.6.3 Contrafactual

Uma explicação contrafactual descreve uma situação causal, oposta ao evento. "Se X não tivesse ocorrido Y, não teria ocorrido". A explicação causal, busca imaginar uma situação hipotética que contradiz os fatos observados (MOLNAR, 2022).

Os métodos contrafactuais, são métodos locais, nas quais, dada uma instância de dados o 'evento' é a previsão da instância local, e as causas são as variáveis inseridas que determinam a previsão. Logo é possível simular eventos contrafactuais, cenários opostos, a previsão de uma instância e entender como as variáveis podem alterar a previsão.

Existem diferentes formas de gerar explicações contrafactuais, a abordagem mais simples é a partir da definição de uma função de perda que avalia a diferença da previsão do modelo e o resultado desejado, essa diferença pode ser minimizada através de uma função de perda que utiliza a distância, por exemplo, a distância de Manhattan como proposta por (WACHTER; MITTELSTADT; RUSSELL, 2018).

Wachter, Mittelstadt e Russell (2018) define uma função de perda através da distância de Manhattan ponderada pelo desvio absoluto mediano (MAD).

$$d(x, x') = \frac{1}{p}\sum_{j=1}^{p}\frac{|x_j - x'_j|}{\text{MAD}_j}$$

Na qual, a distância total é a soma de todas os atributos p, ou seja, as diferenças absolutas dos valores das características entre a instância x e o contra-factual x'. As distâncias por característica são escaladas pelo inverso da mediana da desvio absoluto das características j sobre o conjunto de dados, definido como:

$$\text{MAD}_j = \text{median}_{i \in \{1,\ldots,n\}} \left(\left| x_{i,j} - \text{median}_{l \in \{1,\ldots,n\}} \left(x_{l,j} \right) \right| \right)$$

O cálculo do MAD é a mediana de um vetor, onde realiza a separação pela metade de valores maiores e outra metade menores. O MAD equivale à variância de um atributo, mas ao invés, de utilizar a média e realizar a soma das distâncias quadráticas, é utilizado a mediana como métrica.

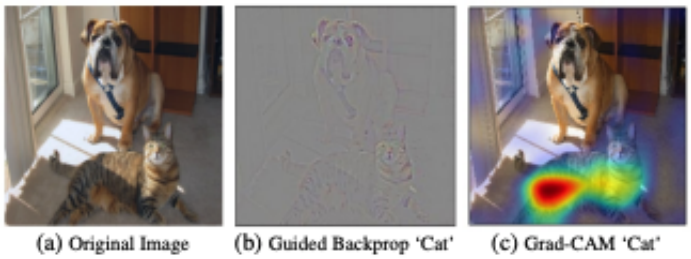
Para minimizar esta função de perda, qualquer algoritmo de otimização adequado pode ser utilizado. A instância x a ser explicada, o resultado desejado y' e o parâmetro de tolerância ϵ devem ser definidos com antecedência. A função de perda é minimizada para x', e o contrafactual x' (localmente) ótimo é retornado enquanto λ é aumentado até que uma solução suficientemente próxima seja encontrada (dentro do parâmetro de tolerância) (MOLNAR, 2022). Para gerar explicações contrafactuais, temos os seguintes passos.

$$\arg\min_{x'} \max_{\lambda} L(x, x', y', \lambda)$$

1. Selecione uma instância x a ser explicada, o resultado desejado y', uma tolerância ϵ e um valor inicial (baixo) para λ.
2. Dada uma instância aleatória como contrafactual inicial.
3. Otimize a perda com o contrafactual inicialmente mostrado como ponto de partida.
4. Enquanto $|\hat{f}(x') - y'| > \epsilon$:
   a) Aumente λ.
   b) Otimize a perda com o contrafactual atual como ponto de partida.
   c) Retorne ao contrafactual que minimiza a perda.
5. Repita os passos 2-4 e retorne a lista de contrafactuais ou aquele que minimize a perda.

A explicação contrafactual gerada pode ser exibida em um formato textual que detalha as diferenças entre os dois cenários. Além disso, a explicação pode ser complementada com imagens que ilustram visualmente essas diferenças, proporcionando uma representação gráfica que ajuda a entender melhor as alterações entre os cenários.

Figura 9 – Método Grad-CAM. Fonte: (SELVARAJU et al., 2019).



2.3.6.4 *Gradient-weighted Class Activation Mapping* - (Grad-CAM)

O Grad-CAM (*Gradient-weighted Class Activation Mapping*) (SELVARAJU et al., 2019) é uma técnica para visualizar as regiões de uma imagem de entrada onde uma Rede Neural Convolucional (CNN) se concentra ao realizar previsões. Após o treinamento e fixação dos parâmetros da rede, o Grad-CAM gera mapas de calor que destacam as áreas da imagem que mais influenciam a decisão do modelo.

Especificamente, o Grad-CAM utiliza os gradientes das classes em relação ao mapa de características das camadas convolucionais finais para identificar quais componentes da imagem têm o maior impacto na pontuação de classificação. Essa abordagem permite não apenas verificar se a rede está olhando para as partes corretas da imagem, mas também facilita a interpretação dos motivos por trás das predições feitas pelo modelo, contribuindo para uma maior transparência e confiança na utilização de redes neurais em aplicações críticas.

(BRUNESE et al., 2020), implementou tal método para explicabilidade em imagens de radiografia relacionados a COVID-19, identificando quais áreas foram mais afetadas dado um paciente com COVID-19, pneumonia e um paciente saudável.

A Figura 9 ilustra a aplicação da técnica Grad-CAM. Nela, a imagem (a) mostra a presença de um cachorro e um gato. Ao realizar a tarefa de classificar o gato, o Grad-CAM gera um mapa de calor que destaca as regiões da imagem com maior importância para essa classificação.

2.4 Métricas de Explicabilidade

As técnicas de explicabilidade visam trazer transparência e entendimento do processo decisório a um modelo de aprendizado de máquina. Apesar de termos diversas técnicas consideradas estado da arte, o que define uma boa explicabilidade?

Para tal, existem as métricas de explicabilidade, que visam medir o quão confiável é determinada explicabilidade. Segundo Doshi-Velez e Kim (2017), podemos dividir as métricas em objetivas, e subjetivas. Segundo Nguyen e Martínez (2020) não é possível obter uma métrica de explicabilidiade que seja aplicável a todas as técnicas de explicabilidade.

Existem algumas dificuldades em se impor uma métrica universal, de maneira similar as métricas de avaliação de aprendizado de máquina, cada métrica de explicabilidade, visa a cobertura de algum aspecto. Além disso, a uma diversidade de métricas de explicação, por exemplo, o LIME aplica pesos a super-pixels, e Grad-CAM atribui a pixels individuais. Métricas únicas dificilmente se aplicam a todas as abordagens. De modo similar o LIME, fornecem explicações de maneiras distintas, desde da própria formulação matemática que rege tais abordagens.

De maneira similar,Vilone e Longo (2020) definem como métricas de explicabilidade podem ser divididas entre métricas orientadas a seres humanos (do inglês, *human centered*) e métricas objetivas. As métricas objetivas podem ser definidas por formalizações e visam entender, por exemplo, a completude da explicação, cardinalidade de regras, métricas de perturbação, etc. A seguir serão discutidas em mais detalhes estas métricas.

2.4.1 Métricas Objetivas

Segundo Kadir e Broberg (2021), em seu estudo, focou em métricas objetivas para explicações locais e identificou métricas objetivas para explicações locais. O foco em explicações locais se dá pelo fato de que explicações locais fornecem *insights* sobre como um modelo de IA chegou a uma determinada previsão ou decisão, permitindo aos usuários entender o raciocínio por trás das saídas do modelo, em relação às instâncias. Algumas métricas definidas por Kadir e Broberg (2021) são: sensibilidade, fidelidade e monotonicidade.

Sensibilidade quantifica quanto uma variável afeta a saída de um modelo. Alguns métodos de explicabilidade já empregam a sensibilidade na própria geração da explicabilidade, como por exemplo, o algoritmo LIME.

Fidelidade é definida como a relação entre as pontuações de importância das características (variáveis), fornecidas por uma função de explicação, e as características realmente relevantes que influenciam a saída do modelo, ou seja, a explicação é considerada fiel se as variáveis destacadas como importantes realmente afetam o desempenho do modelo.

Podemos dividir as métricas de fidelidade (do inglês, faithfulness), em duas categorias: as correlações de fidelidade e as estimativas de fidelidade. A correlação de fidelidade refere-se à relação entre as pontuações de importância atribuídas a características de entrada por uma função de explicação e a real influência que essas características têm sobre a saída do modelo preditivo. Ela mede o quão bem as pontuações de importância se correlacionam com a variação na saída do modelo quando as características são ajustadas.

A estimativa de fidelidade inclui analisar numericamente a correlação entre as pontuações de importância e as saídas do modelo, quanto maior a correlação entre a soma das importâncias a mudança na saída ao modificar o modelo, maior a fidelidade (DASGUPTA; FROST; MOSHKOVITZ, 2022).

Monotonicidade A Métrica de Monotonicidade, introduzida por Luss et al. (2019), gera explicações contrastivas com funções de atributo monotônicas, na qual, refere à propriedade de que uma característica de entrada aumenta (ou diminui), e a importância atribuída a esta variável deve se manter consistente conforme o esperado. A monotonicidade em explicações implica que, para um modelo que deve ser monotônico em relação a uma característica, um aumento nessa característica não deve resultar em uma diminuição na previsão do modelo, e vice-versa.

2.4.2 Métricas Subjetivas

As métricas subjetivas são projetadas para avaliar aspectos centrados no usuário sobre as explicações geradas pelos métodos de explicabilidade. Essas métricas podem ser tanto qualitativas quanto subjetivas e visam capturar a experiência do usuário através das suas percepções, conhecimento a priori sobre o tema e as interações do usuário com as explicações.

No entanto, essas métricas são denominadas subjetivas, pois, cada usuário pode considerar um método mais adequado do que o outro, não avaliando apenas em si a qualidade da explicação, mas também pode estar correlacionada a outros fatores como a experiência do usuário. Este tipo de avaliação pode ser feito por meio de questionários, coleta de opniões etc (NAUTA et al., 2023).

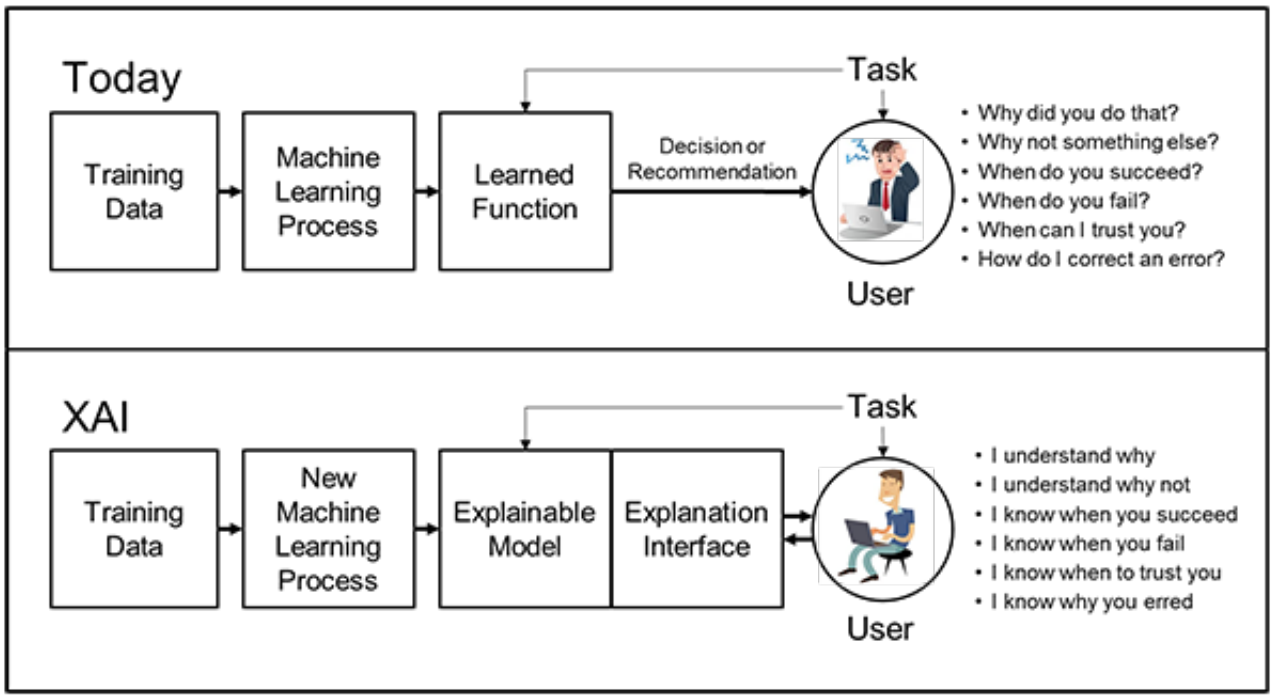
2.5 Explicabilidade para quem?

Em 2023 foi implementado a primeira lei de legislação em relação a Inteligência artificial no mundo na união europeia o, EU AI ACT Europeia (2024), que possui como objetivo regulamentar as aplicações de inteligência artificial com base no seu risco em causar prejuízos, divididos em quatro categorias: inaceitáveis, riscos altos, limitados e mínimos. Sendo níveis inaceitáveis banidos e de alto risco, métodos que requerem transparência como obrigatório.

Segundo Jung et al. (2023), a explicabilidade pode ser vista como um requisito para a confiabilidade em IA, visto que a forma com que os modelos representam a incerteza influencia na compreensão e na tomada de decisão dos usuários, logo sendo necessário a explicabilidade para poder mitigar tais problemas.

Além de tal fato, atualmente temos produtos denominadas "*AI-First*"(FONTANA, 2021), que utilizam a AI, como elemento fundamental, integrando em todos os aspectos do produto, desde a concepção até a integração, fornecendo ao usuário experiência personalizadas, como por exemplo, *Vision Pro* da Apple, que está sendo desenvolvido por meio da integração da IA a inúmeras ferramentas.

Figura 10 – Integração da explicabilidade em produtos de AI Gunning D.; Vorm e Turek (2021).



Diante dessa necessidade por parte das empresas na adoção da IA em toda sua cadeia de desenvolvimento. Segundo Gunning D.; Vorm e Turek (2021), a explicabilidade será integrada juntamente com o desenvolvimento dos próprios modelos, ou seja, tais sistemas de aprendizado de máquina trariam consigo explicações intrínsecas, como observado na figura abaixo.

A figura 10, exibe a integração da XAI no desenvolvimento de tais produtos. Através de uma ferramenta de explicação o usuário teria uma confiança sobre o que o modelo que está sendo utilizado. Desta forma, aprimora a assertividade quanto a sua aplicação prática.

2.6 Sistemas de Suporte à Decisão Clínica

Sistemas de Suporte à Decisão Clínica (SSDC) são plataformas que utilizam dados específicos de cada paciente e conhecimento embasado em evidências para oferecer orientações, avaliações ou sugestões personalizadas aos profissionais de saúde. Esses sistemas atuam como auxiliares, capacitando os clínicos a tomar decisões mais fundamentadas e precisas quanto ao tratamento mais adequado para cada paciente em particular. Os SSDC podem ser integrados aos registros médicos eletrônicos (EMR), fazendo uso de algoritmos sofisticados e análises de desfechos para prover apoio decisório baseado em evidências (DINEVSKI et al., 2011).

As implementações dos SSDC são caracterizadas por diversos elementos. Eles fazem uso de métodos de inferência, algoritmos ou técnicas para processar informações. Utilizando detalhes do paciente, como histórico médico, resultados de exames e outros dados pertinentes, esses sistemas fornecem recomendações customizadas. O resultado é direcionado a um profissional humano, que pode ser um médico, enfermeiro ou outro especialista da saúde necessitando de orientação para decisões clínicas embasadas e precisas (GREENES, 2011). A inferência realizada busca oferecer recomendações individualizadas e respaldadas por evidências aos profissionais, com o propósito de aprimorar a qualidade do cuidado prestado e minimizar erros médicos.

2.7 Conhecimento como Serviço

A partir dos SSDC, surge a necessidade de arquiteturas de referência para determinado domínio de aplicação, tais arquiteturas são definidas como modelos generalistas, que definem componentes fundamentais e suas relações (CLEMENTS, 2002). As arquiteturas orientadas à serviço (*Service-oriented architecture* - SOA), segundo McGovern et al. (2006), são definidas como um estilo arquitetônico baseado em componentes, na quais os módulos provêm serviços a outros módulos, permitindo a integração de diferentes serviços tanto internos quanto externos à organização.

Originado do modelo SOA, surgiram outros paradigmas, incluindo o software como serviço (SaaS) e a arquitetura baseada no conhecimento como serviço (KaaS). Esta última busca centralizar e fornecer conhecimento, extraído de diversas fontes de dados, por meio de serviços bem definidos, permitindo acesso facilitado para os consumidores.

Segundo Xu e Zhang (2005), o paradigma de conhecimento como serviço (do inglês, *knowledge as a service*), é uma abordagem que visa fornecer o conhecimento ao usuário (consumidor do conhecimento), através de um serviço, sendo possível encontrar três componentes principais: *Data Owners* (Proprietários de Dados), *Knowledge Extractor* (Extrator de Conhecimento) e *Knowledge Server* (Servidor de Conhecimento).

Os Proprietários de dados, são as entidades que possuem e mantêm os dados que são relevantes para a extração do conhecimento, são responsáveis por proteger a privacidade e a segurança dos dados, ao mesmo tempo que permite que o conhecimento seja extraído através de serviços para outras entidades.

Extrator de conhecimento possui como sua responsabilidade extrair o conhecimento a partir do conjunto de dados fornecido através dos *Data Owners*, o extrator se utiliza de técnicas de mineração de dados e/ou algoritmos de aprendizado de máquina para analisar e/ou criar modelos.

Servidor de conhecimento visa fornecer o serviço de conhecimento aos consumidores, através dos modelos de conhecimento extraídos pelo extrator de conhecimento para

Figura 11 – Arquitetura conceitual H-KaaS (BARRETO et al., 2018).

responder às consultas realizadas pelos consumidores. Atuando como uma ponte entre os dados dos proprietários e os consumidores de conhecimento.

2.7.1 H-KaaS (Health-Knowledge as a Service)

A partir da ideia de um serviço centralizado (BARRETO et al., 2018), propõe o H-KaaS (*Health Knowledge as a Service*), um serviço de KaaS dedicado ao domínio da saúde, adaptado a partir da definição de (XU; ZHANG, 2005), que define os principais componentes da arquitetura dedicada a saúde.

A arquitetura H-KaaS, possui os mesmos três componentes essenciais da arquitetura KaaS, exibidos pela Figura 11. Podemos considerar como fontes de dados, fontes que ainda não alimentaram processos de aprendizado de máquina, como bases de dados de domínio médico, prontuários médicos eletrônicos, planilhas, imagens, diretrizes clínicas etc.

Os consumidores acessam o servidor por meio de um protocolo autenticado e bidirecional de troca de mensagens, através de uma API (*Application Programming Interface*). Através das requisições o conhecimento é solicitado ao servidor do conhecimento, que então responde com base nas conclusões derivadas pelo raciocinador a partir do modelo de dados.

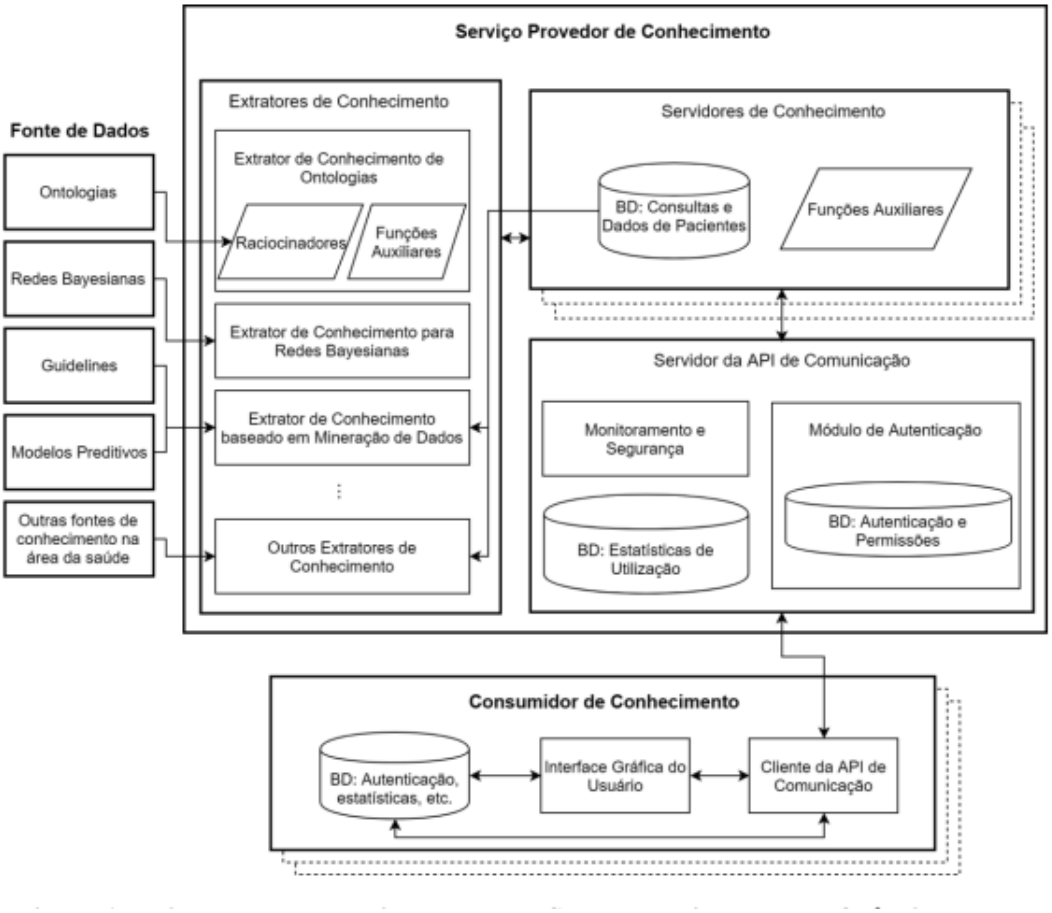
A figura 12 exibe a arquitetura detalhada do H-KaaS e seus respectivos módulos, que serão detalhados na seção a seguir.

2.7.1.1 Servidor provedor de conhecimento

Dentro do paradigma KaaS, o módulo provedor de conhecimento tem a finalidade de acessar e processar dados da fonte de dados, gerenciar modelos de conhecimento e atender consultas dos aplicativos consumidores de conhecimento.

Para atender às consultas dos consumidores, o servidor implementa uma API de comunicação que fornece respostas padronizadas, de fácil compreensão para os aplicativos

Figura 12 – Arquitetura detalhada H-KaaS (BARRETO et al., 2018).



consumidores de conhecimento. A extração de conhecimento pode ocorrer de diversas maneiras no contexto dos aplicativos consumidores. Isso pode envolver inferências na ontologia, consultas a documentos ou outras fontes de dados e conhecimento. Alternativamente, o conhecimento pode ser extraído periodicamente de bancos de dados usando algoritmos de aprendizagem de máquina e KDD (*Knowledge Discovery in Databases*).

Cada fonte de dados possui suas próprias regras de extração, que são implementadas nos extratores de conhecimento. Assim, sempre que uma nova fonte de dados é adicionada, é necessário escrever regras de extração e acesso para integrá-la ao sistema existente.

Os extratores de conhecimento também incluem algoritmos auxiliares para várias funções comuns, como leitura de arquivos de texto, manipulação de imagens e aplicação de filtros para garantir o anonimato dos dados. Além disso, os extratores podem utilizar outros mecanismos de persistência presentes nos serviços provedores de conhecimento para aprimorar o conhecimento modelado.

2.7.1.2 Consumidor de Conhecimento

Na arquitetura H-KaaS, semelhante ao paradigma KaaS, é possível que diferentes aplicativos consumidores de conhecimento acessem o sistema através da API de comunica-

ção para consultar uma ou mais bases de conhecimento.

A implementação de cada aplicativo pode variar de acordo com seus objetivos e as linguagens de programação escolhidas, contanto que a comunicação com a API siga suas especificações.

Em termos de segurança, cada aplicativo deve ter uma chave única e privada para autenticar, limitar e/ou identificar as consultas realizadas ao serviço provedor. A falta dessa chave de segurança impedirá o acesso ao serviço, sendo a responsabilidade do serviço provedor criar, armazenar e compartilhar essas chaves.

Entre os possíveis aplicativos consumidores, no domínio da saúde, podemos mencionar *websites* para apoio à decisão clínica, aplicativos móveis ou embarcados, prontuários eletrônicos, sistemas educacionais, entre outros. Quanto aos serviços oferecidos, há um comando que recebe requisições do tipo GET e lista os serviços atualmente disponíveis na plataforma. Cada serviço inclui descrições, códigos de identificação e, principalmente, uma lista de métodos disponíveis.

Os métodos são funções que podem ser executadas a qualquer momento pelo aplicativo consumidor, permitindo o acesso ao conhecimento disponível. A documentação de cada objeto está disponível para consulta. O comando responsável pela execução de um método específico recebe requisições no formato POST e espera que o aplicativo forneça informações serializadas com base no formulário de entrada presente na chave de entrada. A resposta a esse comando é um objeto serializado em JSON, que inclui informações sobre a consulta realizada e o resultado da execução (BARRETO et al., 2018).

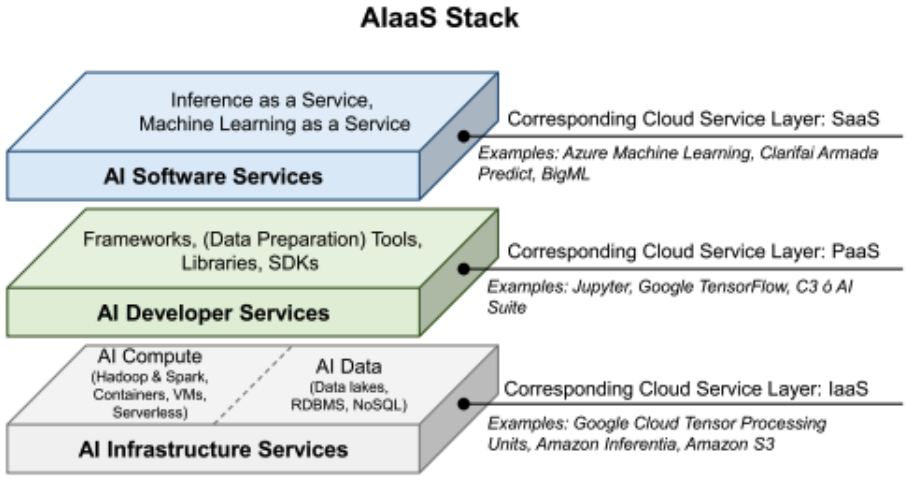
2.8 Inteligência Artificial como Serviço

Visando estabelecer um padrão para desenvolvimento de modelos de aprendizado, aliados às práticas mais recentes de desenvolvimento de software, surge a Inteligência Artificial como Serviço (do inglês, Artificial Intelligence as a Service) (AIaaS), Lins et al. (2021). Visa garantir não apenas a criação de modelos de aprendizado de máquina eficazes, mais também o acesso sob demanda a um conjunto compartilhado de recursos de computação configuráveis que possam ser rapidamente provisionados.

AIaaS tem como objetivo tornar a IA acessível para todos, independente do tamanho da organização, conhecimento tecnológico e orçamento disponível. O paradigma AIaaS orienta seus usuários no processo de desenvolvimento, implantação ou uso de modelos de análise de dados sem a necessidade de aprender algoritmos ou tecnologias complexas Elshawi e Sakr (2017). AIaaS pode ser dividido em três camadas:

- **Serviços de software de IA**: São aplicativos de IA prontos para uso e blocos de construção (relacionados à camada de nuvem SaaS convencional), também referidos como Inferência como serviço, onde os usuários podem acessar modelos de aprendizado

Figura 13 – Visão Conceitual da Arquitetura AIaaS (LINS et al., 2021).



de máquina pré-treinados ou modelos de aprendizado de máquina como serviço (*Machine Learning as a Service* - MLaaS), permitindo a customização dos modelos. Essa camada de inferência pode ser acessada via API.

- **Serviços de desenvolvimento de IA**: Ferramentas para ajudar os desenvolvedores a implementar código para explorar as capacidades de IA.
- **Serviços de infraestrutura de IA**: que compreendem potência computacional bruta para construir e treinar algoritmos de IA, e capacidades de rede e armazenamento para armazenar e compartilhar dados (relacionados à camada de nuvem de *Infrastructure as a Service* - IaaS convencional).

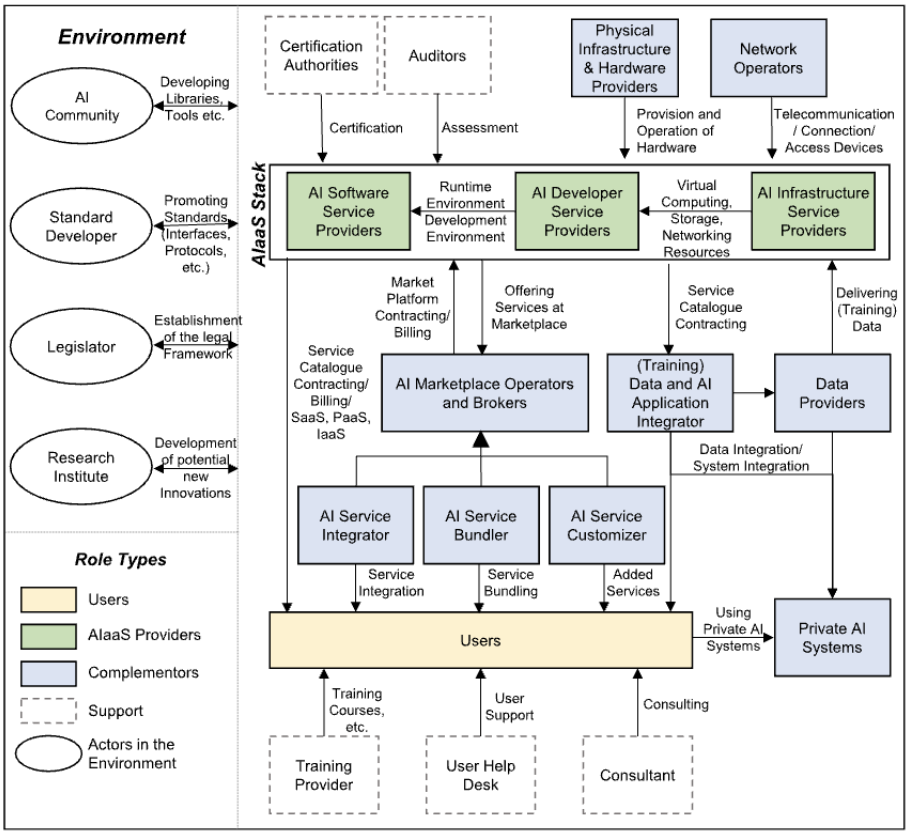
Essa estrutura é viável através dos seus componentes, pois ao utilizar a infraestrutura da nuvem, temos um arsenal de mecanismos que garantem a robustez da aplicação, oferecendo desempenho, e garantindo alternativas caso o sistema apresenta falhas.

A Figura 13 oferece uma visão conceitual, em camadas da arquitetura de AIaaS e suas principais definições. Apesar da imagem exemplificar três camadas, as camadas ***AI Developer Services*** e ***AI Infrastructure Services***, não serão detalhadas, visto que a camada ***AI Developer Service***, é referente a utilização de biblioteca e ferramentas, que podem ser disponibilizadas pelos serviços na nuvem, como ferramentas de armazenamento de dados e disponibilização de máquinas virtuais.

A camada ***AI Infrastructure Services***, é referente aos serviços de infraestrutura que as plataformas na nuvem fornecem, não sabemos como tais serviços foram desenvolvidos, apenas utilizamos em alguns casos instâncias de tais serviços de infraestrutura apenas em alto nível. Exemplo destes serviços são ferramentas de segurança.

A Figura 14 ilustra o ecossistema completo do AIaaS, destacando os diversos papéis envolvidos na arquitetura. Nosso foco será no ***AIaaS Stack***, mais precisamente no

Figura 14 – Arquitetura detalhada AIaaS (LINS et al., 2021).



módulo, *AI Software Services*, que serão detalhados a seguir.

2.8.1 *AI Software Services*

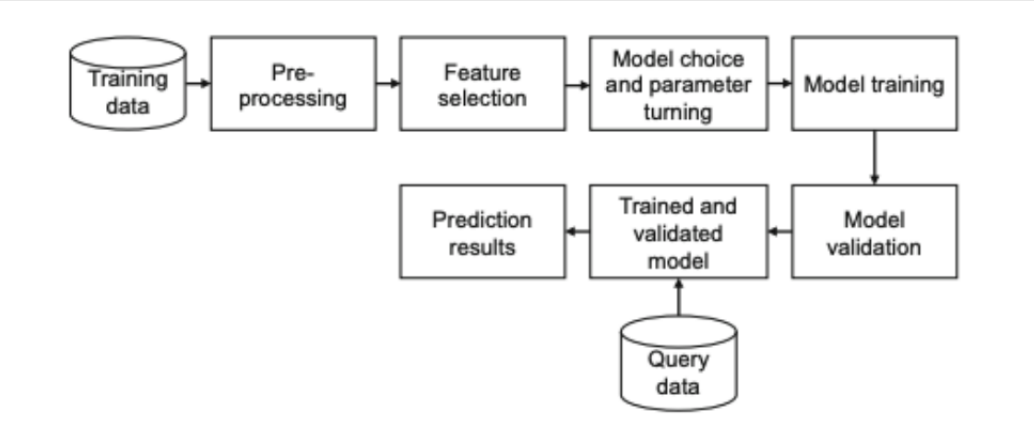
O módulo de serviço de software de IA refere-se à construção de aplicações que utilizam inteligência artificial e as disponibilizam para uso sob demanda. Segundo Lins et al. (2021), tal módulo também pode ser denominado como *Machine Learning as a Service* (MLaaS) ou *Inference as a Service*, sendo este último referente às inferências realizadas por modelos de aprendizado de máquina

No entanto, há uma lacuna na definição desta camada, uma vez que o conceito de IA é mais amplo do que apenas aprendizado de máquina, incluindo também inferências lógicas e o uso de agentes.

AI Software Services é responsável por gerenciar desde a etapa de tratamento de dados até a disponibilização do modelo já treinado, que pode ser acessado via API. Embora não sejam especificados os componentes essenciais dessa camada, existem algumas arquiteturas descritas na literatura, como a proposta por Yao et al. (2017), Figura 15. Nesta abordagem, é apresentada uma arquitetura de MLaaS, que inclui um arquitetura responsável por coletar, depurar e transformar os dados para o treinamento do modelo.

De modo similar a figura proposta por Fayyad (2001), a arquitetura de Yao et al.

Figura 15 – Arquitetura de MlaaS (YAO et al., 2017).



(2017), exemplifica passos para o processo de modelagem de aprendizado de máquina, que incluem etapas de reunir informações de múltiplas fontes, corrigir inconsistências ou omissões nos dados e formatá-los de maneira compreensível para o modelo. Além disso, a pipeline pode envolver a combinação de dados e criação de novas variáveis por meio de técnicas de *Feature Engineering*.

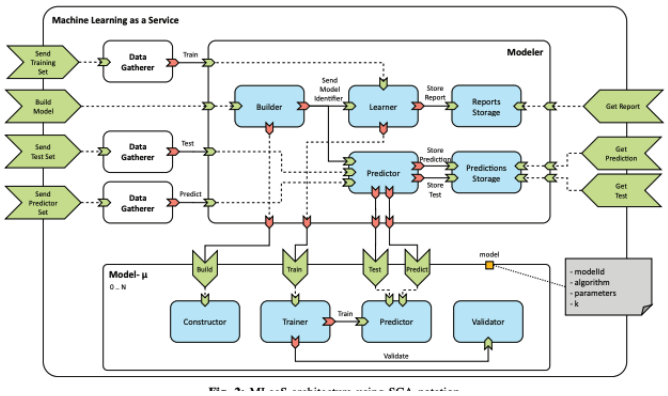
De modo similar Ribeiro, Grolinger e Capretz (2015), propõe uma abordagem flexível e escalável, baseada na *Service Component Architecture (SCA*), na qual, é uma especificação de modelagem que visa compor sistemas com os princípios da SOA. A arquitetura proposta, é projetada para suportar o aprendizado de máquina, a partir de múltiplas fontes de e construção de vários modelos utilizando diferentes algoritmos, como podemos visualizar na figura Figura 16.

Além do treinamento, ambas arquiteturas se encarregam de ajustar hiperparâmetros, validar o modelo e disponibilizá-lo, assim sendo possível consumir o modelo via requisições. A partir de tais componentes é possível oferecer uma infraestrutura robusta ao usuário final.

De modo geral, a camada de *AI Software Services*, é responsável pela integração de dados nas plataformas da nuvem, a preparação de dados, que engloba os passos como a normalização dos dados, a engenharia de características ou o aumento de imagens, bem como a divisão dos conjuntos de treinamento e teste.

Com os dados devidamente preparados, inicia-se o processo de treinamento do modelo. Escolhe-se um algoritmo de aprendizado de máquina específico e são definidos seus parâmetros e hiperparâmetros. A seguir, ocorre a avaliação do modelo, na qual um conjunto de métricas de avaliação é estabelecido para mensurar o desempenho. Após a avaliação das métricas de desempenho, o modelo escolhido será integrado e disponibilizado para uso dos usuários.

Figura 16 – Arquitetura de MlaaS proposta por (RIBEIRO; GROLINGER; CAPRETZ, 2015).



2.8.2 API de Comunicação e Microsserviços

A comunicação entre a arquitetura proposta e o aplicativo consumidor é feita por meio de uma API.

Uma API é um conjunto de recursos e funcionalidades que permitem que aplicativos diferentes se comuniquem e compartilhem dados. Uma de suas principais vantagens é o encapsulamento, que permite que a lógica da API seja ocultada do aplicativo consumidor. Nos anos 2000, um padrão de arquitetura para APIs foi proposto por FIELDING R.T; RICHARD (2000b), denominado de *Representational State Transfer* (REST), método que se tornou um padrão para comunicação entre cliente/servidor, na qual os mesmos são independentes e as requisições sempre são feitas pelo cliente, um outro ponto de destaque para tal arquitetura é que cada requisição é independente, não existe dependência de uma requisição passada. Tal arquitetura utiliza-se de protocolos HTTP (H*ypertext Transfer Protocol*), protocolo que se tornou padrão para comunicação na internet.

Uma API RESTful, utiliza de cinco comandos existentes no protocolo HTTP, definidas pelo protocolo RFC 2616, cada comando é responsável por uma operação, tais métodos são: GET, POST, PUT, DELETE e PATCH. O Comando POST, é responsável pelo envio de dados ao servidor (criação de um novo recurso), para aplicações de aprendizado de máquina que é utilizado pelo usuário para inferir predições. O Comando GET é utilizado para ler ou recuperar dados de um servidor. O comando DELETE, pode remover uma informação no servidor, como a remoção de dados. Já os comandos PUT e PATCH, são utilizados para atualizar ou modificar um dado.

A comunicação entre as requisições e os resultados de tais requisições, é por um dado semiestruturado para serialização em formato textual legível por humanos para armazenar e transmitir dados compostos por pares de atributos e valores, denotado de *JavaScript Object Notation* (JSON).

Será utilizado a comunicação via API para a nossa arquitetura, utilizando o formato

JSON para comunicação, baseada na arquitetura REST. A API projetada terá apenas requisições no formato POST, o comando é responsável pela execução de dois métodos, o formulário de envio de dados para a execução da predição e um segundo formulário com o *feedback* da predição.

Ao receber requisições, o aplicativo fornecerá informações baseadas no formulário, em seguida tais respostas computadas serão serializadas em um objeto JSON. O tratamento de erros, a aplicação deverá ter um objeto denominado ExceptionError, que será instanciado quando algum problema ocorre na aplicação, possíveis erros são: erros de autenticação, formato inválido de dados, serviço indisponível, etc.

Visando a segurança do nosso sistema, a execução de cada chamada do (API), é necessário chaves para comunicação, garantindo acesso aos serviços do modelo. O processo de autenticação é realizado pela plataforma, garantindo o controle da plataforma.

3 Trabalhos Relacionados

Com o fácil acesso a grandes volumes de dados, as empresas encontram maior facilidade para adotar técnicas de aprendizado de máquina e compreender padrões em suas informações, as técnicas de explicabilidade em conjunto com esses modelos, ajudam a preencher lacunas entre a complexidade do modelo e tornar mais transparente e claro o processo da tomada de decisão.

É necessário que tanto usuários leigos quanto especialistas consigam entender e confiar nos resultados gerados por tais modelos com clareza. Em domínios mais críticos, como a área da saúde, a explicabilidade desempenha um papel fundamental. A capacidade de interpretar as decisões dos modelos de aprendizado de máquina não só aumenta a confiança dos usuários, mas também garante a transparência e a responsabilidade no uso dessas tecnologias.

Além disso, a explicabilidade atua como um catalisador para a implementação de medidas regulatórias. Dado o crescente número de leis e regulamentações sobre o uso de modelos de aprendizado de máquina, a capacidade de explicar e interpretar os resultados destes modelos torna-se essencial para garantir a conformidade.

No entanto, a integração da explicabilidade em aplicações reais ainda é uma lacuna e cheia de desafios, tanto técnicos quanto pessoais, seja da adaptação dos usuários a interagir com as explicações fornecidas quanto o impacto econômico para implementação.

Além disso, não existe uma arquitetura de referência, para a integração de explicabilidade a modelos de aprendizado de máquina tradicionais.

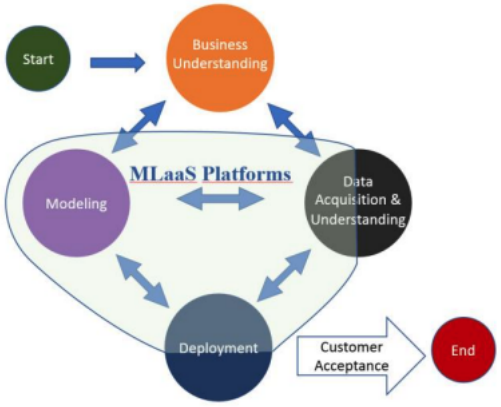
Nesta revisão narrativa, buscamos analisar diversos estudos, *frameworks* e *pipelines* de aprendizado de máquina em conjunto com técnicas de explicabilidade, bem como os módulos associados a essas abordagens. Além disso, exploramos o impacto dessas arquiteturas tanto do ponto de vista acadêmico quanto do ponto de vista de negócios.

3.1 Arquiteturas para Aprendizado de Máquina

Philipp et al. (2021), no trabalho proposto intitulado de "***Machine Learning as a Service – Challenges in Research and Applications***", elenca alguns desafios relacionados ao temas de disponibilização de serviços de aprendizado de máquina presentes na literatura, entre os temas debatidos, são as aplicações.

De modo geral, o artigo de Philipp et al. (2021) apresenta um ciclo de vida de desenvolvimento de produtos de aprendizado de máquina, podemos dividir em três principais etapas, como vistos na Figura 17. Aquisição de dados, tal fato se refere a

Figura 17 – MLaaS no contexto de Ciência de Dados Fonte: (PHILIPP et al., 2021)).



dados que não estão conectados ao ambiente de desenvolvimento, além de etapas como manipulação, exploração e limpeza dos dados que são uma processos frequentemente utilizados no desenvolvimento.

Em seguida temos a etapa de modelagem, ferramenta essencial para plataformas de MLaaS e consequentemente AIaaS, que é a construção do modelo. Por fim, a disponibilização do modelo, para que possa ser consumido pelos usuários, através de uma API.

De modo Lewicki et al. (2023), define em seu artigo "***Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service***", elaborar uma taxonomia em AIaaS e elenca os serviços em três categorias: Aplicações de AutoML (Aprendizado automatizado), aplicações via API e por fim, os serviços de IA totalmente gerenciados.

As aplicações de AutoML, visa a construção de modelos de aprendizado de máquina com pouca intervenção humana, onde certas etapas do desenvolvimento são automatizadas. Desde de seleção de hiperparâmetros, modelos e seleção de variáveis.

Uma das grandes vantagens dessa abordagem é a criação de modelos com poucos "cliques". No entanto, ela apresenta algumas limitações, desde problemas clássicos na modelagem até questões inerentes à plataforma, como os custos. Como esses serviços funcionam de forma exaustiva, o orçamento se torna um delimitador crucial para encontrar o melhor modelo para o usuário.

As arquiteturas AIaaS que oferecem serviços via API, disponibiliza modelos de aprendizado de máquina pré-construídos através de serviços, nas quais, os usuários podem consumi-los através de uma solicitação via API. Tais modelos atuam em diversos segmentos, desde dados tabulares para tarefas mais simples como uma classificação a tarefas mais complexas como serviços de visão computacional.

De modo similar, provedores na nuvem também oferecem serviços que é possível o usuário criar o próprio o modelo e disponibilizá-lo online, sendo possível acessar através

de uma requisição via API.

Os serviços de IA totalmente gerenciados, oferecem uma visão completa de uma arquitetura para o usuário, sendo capazes de gerenciar múltiplos recursos do usuário, como múltiplas fontes de dados e interfaces, atendendo assim a diferentes necessidades da indústria (LEWICKI et al., 2023).

Markov et al (MARKOV et al., 2022), propõem em seu artigo "***Looper: an end-to-end ML platform for product decisions***", uma plataforma de aprendizado de máquina, de ponta a ponta projetada para ajudar engenheiros de produto a tomar decisões orientadas por dados com facilidade, na arquitetura proposta, inclui a coleta de dados em tempo real, treinamento de modelos, avaliação de modelos, implantação de modelos em produção e monitoramento contínuo do desempenho do modelo.

As pipeline de treinamento e inferência no Looper, a plataforma inclui ferramentas de monitoramento que rastreiam o uso de recursos nos componentes do pipeline de treinamento e inferência. De modo geral, o Looper é projetado para fornecer pipelines de treinamento e inferência eficientes e eficazes para modelos de aprendizado de máquina. A comunicação é realizada via API, para o testes A/B e consumo de maneira geral.

(KHATTAK et al., 2023), no artigo "***MLHOps: Machine Learning for Healthcare Operations***", fornece uma visão geral do trabalho nesta área e diretrizes para desenvolvedores e clínicos implantarem e manterem seus próprios modelos na prática clínica.

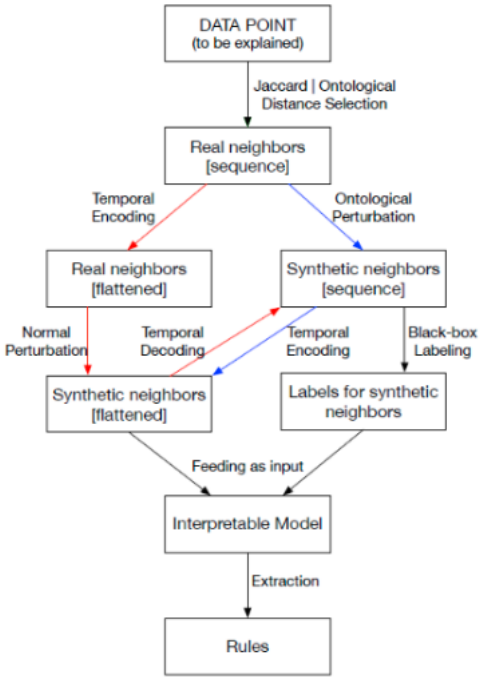
O artigo aborda os diferentes componentes das pipelines de **MLHOps**, incluindo fontes de dados, preparação, engenharia e ferramentas. De maneira semelhante aos artigos expostos anteriormente, possui na sua arquitetura a priorização da privacidade e segurança dos dados do paciente, bem como a interpretabilidade dos modelos de aprendizado de máquina. Além disso, a arquitetura suporta o monitoramento contínuo dos modelos em produção, permitindo a detecção e correção de desvios nos dados e nos modelos.

Metta et al. (2024) propõe em seu artigo a ferramenta denominada de DoctorXAI, Figura 18, ferramenta que busca auxiliar clínicos a entender as decisões dos modelos de aprendizado de máquina. Ela utiliza um classificador para problemas de multi-classes, que analisa o histórico clínico de um paciente para prever a próxima visita, gerando explicações interpretáveis a partir de um modelo de árvore de decisão.

O processo envolve a criação de um "vizinhança sintética" a partir de dados reais, permitindo que o modelo extraia explicações baseadas em regras que são relevantes para casos individuais.

A arquitetura do DoctorXAI fornece explicações interpretáveis para decisões de modelos de aprendizado de máquina em contextos clínicos. Primeiro, seleciona dados semelhantes ao caso a ser explicado e gera uma vizinhança sintética perturbando esses dados, tal processo exemplificado na figura 18.

Figura 18 – DoctorXAI *pipeline*. Fonte: (METTA et al., 2024)



Em seguida, rotula essas instâncias sintéticas com um modelo *black-box* e treina um modelo interpretável, como uma árvore de decisão. Finalmente, extrai explicações baseadas em regras, detalhando como os atributos dos dados influenciam as decisões. Isso traduz padrões complexos de IA em informações compreensíveis, facilitando a colaboração e a compreensão na área da saúde.

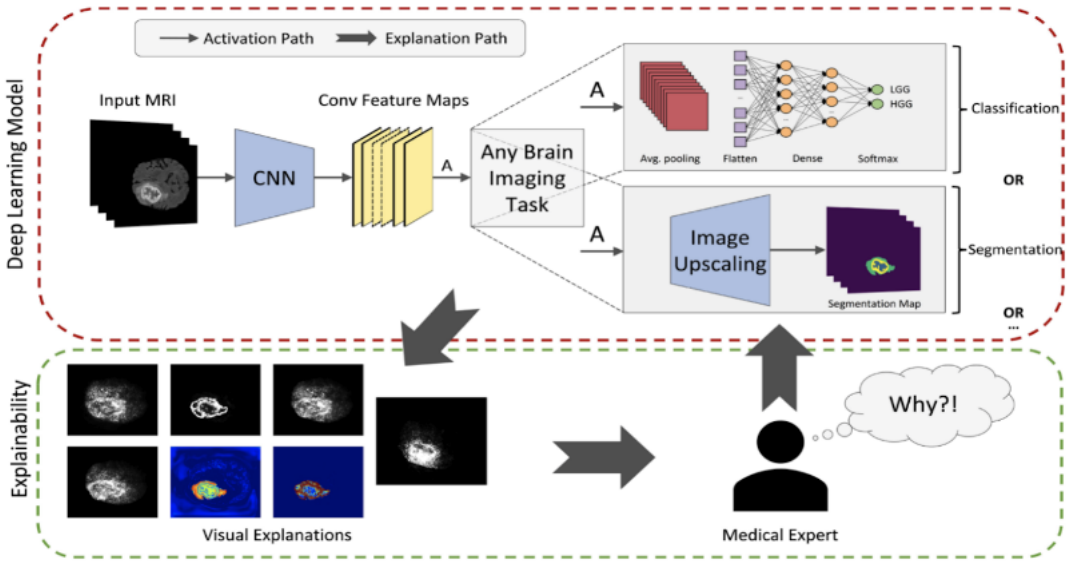
Zeineldin R.A. (2022), em seu artigo "***Explainability of deep neural networks for MRI analysis of brain tumors***", aborda uma arquitetura denominada de Neuro-XAI, *framework*, Figura 19, desenvolvido para aumentar a transparência em modelos de redes neurais profundas relacionados à imagens de ressonância magnética.

A arquitetura é dividida em dois módulos, ***Deep Learning Model*** e ***Explainability***. O primeiro módulo é responsável pelo processamento das imagens e geração das tarefas de classificação e/ou segmentação. Em seguida temos o módulo que gera as explicabilidades, a partir da saída da rede neural, este módulo fornece visualizações que buscam auxiliar na interpretação dos resultados.

O trabalho apresentado por Wang et al. (2024) propõe uma estrutura de serviços para a integração de técnicas de explicação de modelos de inteligência artificial (XAI) em ambientes de *Machine Learning Operation*s (MLOps), focando na gestão do ciclo de vida de desenvolvimento de modelos de aprendizado de máquina.

Denominada XAIport, essa estrutura é composta por quatro componentes principais: o centro de coordenação, o processamento e armazenamento de dados, os microsserviços de XAI e os microsserviços de avaliação.

Figura 19 – NeuroXAI *pipeline*. Fonte: (ZEINELDIN R.A., 2022)



O centro de coordenação desempenha o papel de gestão da estrutura, supervisionando desde a entrada de dados até a geração das explicações. O componente de processamento e armazenamento de dados é responsável pelo gerenciamento dos dados de entrada e das explicações geradas. Os microsserviços de XAI armazenam os métodos de XAI e calculam as contribuições dos modelos de IA. Por fim, os microsserviços de avaliação se dedicam à mensuração das métricas de explicabilidade.

3.2 Abordagens de XAI para dados tabulares

O estudo proposto por Okay, Yıldırım e Özdemir (2021), intitulado de "***Interpretable Machine Learning: A Case Study of Healthcare***", demonstra em seu estudo a importância das aplicações de técnicas de explicabilidade no contexto da saúde. Isto ocorre, devido que os recentes avanços e o amplo uso de técnicas "caixa-preta"afetam o entendimento do modelo. Logo, em aplicações críticas como diagnóstico médico, é fundamental entender os motivos por trás das decisões dos modelos.

No estudo, foram empregados modelos de *boosting*, tais como os algoritmos *Random Forest* e *Gradient Boosting*, juntamente com as técnicas de explicabilidade LIME e SHAP, amplamente utilizadas para lidar com conjuntos de dados tabulares. Um dos principais destaques do estudo foi que, embora tenham sido utilizadas técnicas complexas de aprendizado de máquina, o uso das ferramentas de explicabilidade permitiu compreender o processo decisório do modelo, mantendo, assim, sua acurácia, como demostrado na figura 20.

O estudo realizado por (RASHED-AL-MAHFUZ et al., 2021), "*Clinically Applicable*

Figura 20 – Resultados do LIME. *pipeline*. Fonte: (OKAY; YıLDıRıM; ÖZDEMIR, 2021)

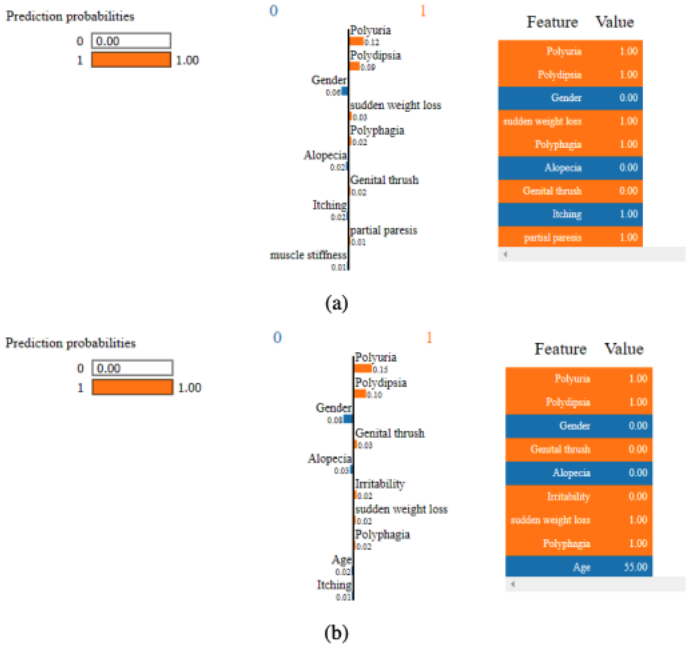
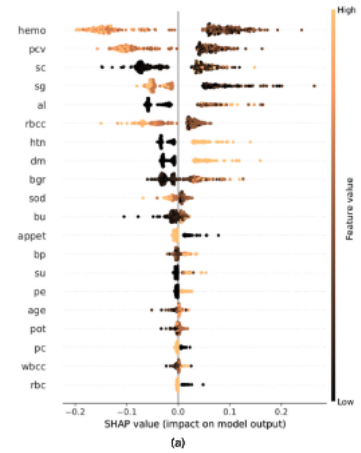


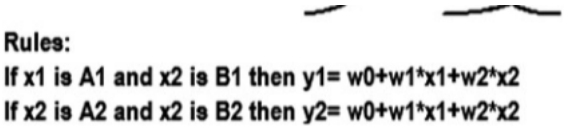
Figura 21 – Resultados do estudo proposto por (RASHED-AL-MAHFUZ et al., 2021).



Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening", investiga o uso de aprendizado de máquina para identificar atributos relevantes na detecção de doença renal crônica (DRC). A pesquisa se baseia em um conjunto de dados clínicos que inclui variáveis como idade, pressão arterial e níveis de creatinina, entre outros indicadores.

Para desenvolver o modelo de aprendizado de máquina, foram testados vários algoritmos, incluindo *Random Forest*, *Gradient Boosting* e Máquinas de Vetores de Suporte (SVM). Além disso, a técnica de explicabilidade SHAP foi empregada para identificar os atributos que mais contribuíram para a classificação da condição, a Figura 21, exibe o resultado com as aplicações de tais técnicas.

Figura 22 – Representação das Regras Fuzzy geradas. Fonte: (SETTOUTI; CHIKH; SAIDI, 2012)



No artigo *Machine learning-based prediction of in-hospital mortality using admission laboratory data: A retrospective, single-site study using electronic health record data.*"(SEKI; KAWAZOE; OHE, 2021), os autores exploram a construção de um modelo de aprendizado de máquina para prever a mortalidade hospitalar dentro de 14 dias após a admissão. A previsão da mortalidade é vista como essencial para otimizar a alocação de recursos de acordo com a necessidade de cada paciente.

Para a previsão de mortalidade, foram utilizadas 25 variáveis, incluindo idade, sexo e atributos laboratoriais. Com o objetivo de identificar o melhor modelo para essa tarefa, foram testados quatro modelos: Regressão Logística, *Random Forest*, Árvores de Decisão e uma Rede Neural.

O estudo evidenciou a eficácia de ferramentas preditivas, destacando seu potencial como instrumento valioso para avaliar o risco de mortalidade em pacientes hospitalizados. A técnica SHAP foi aplicada para gerar explicações sobre as decisões do modelo, e o artigo incluiu o *feedback* de especialistas quanto à qualidade das explicações, garantindo que os *insights* fornecidos fossem úteis para a tomada de decisões clínicas.

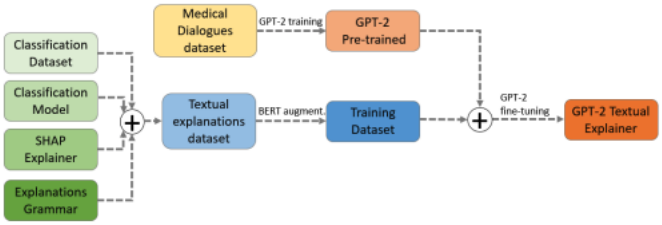
No estudo de Settouti, Chikh e Saidi (2012), os autores exploram o desenvolvimento de uma classificação fuzzy para a detecção de diabetes, utilizando o algoritmo Fuzzy C-Means em conjunto com um sistema de inferência neuro-fuzzy (ANFIS). O objetivo do algoritmo é identificar casos de diabetes tipo 2.

O ANFIS combina redes neurais e sistemas de inferência fuzzy, integrando as vantagens de ambos os paradigmas: a robustez das classificações obtidas pelas redes neurais e a explicabilidade das inferências fuzzy. Como o modelo se baseia em lógicas fuzzy, a explicabilidade é construída por meio de regras e oferece explicações em linguagem natural, o que facilita a compreensão das decisões pelo viés linguístico. As explicações geradas podem ser visualizadas na figura 22.

A qualidade das explicações geradas pelo modelo foi avaliada em termos de compreensibilidade e interpretabilidade, o que sugere que especialistas ou clínicos poderiam ter sido envolvidos nesse processo. A capacidade de fornecer explicações que são intuitivas e que podem ser entendidas por profissionais de saúde é um aspecto crucial para a aceitação e utilização prática do modelo em ambientes clínicos.

O estudo "*Interpretability of Artificial Hydrocarbon Networks for Breast Cancer Classification*", proposto por Ponce e Martinez-Villaseñor (2017), aborda a aplicação de

Figura 23 – Geração de explicações textuais. Fonte: (TORRI, 2021)



redes hidrocarbonadas artificiais aplicadas ao câncer de mama. Para o estudo foi utilizada uma base de dados pública relacionados ao câncer de mama, tal base continha informações referentes ao paciente quanto a dados clínicos.

As redes hidrocarbonadas artificiais (AHN) são um método de aprendizado de máquina, baseados em estruturas das redes de carbono na química, por ser uma estrutura baseada em carbonos, tal estrutura favorece a interpretabilidade, devido sua representação hierárquica e visual das relações entre as variáveis.

O estudo aponta que aplicações de AHN são eficazes para problemas de classificação e em gerar modelos interpretáveis, sendo uma opção viável para superar a dualidade entre a precisão e interpretabilidade que caracteriza muitos modelos de aprendizado de máquina.

No estudo foi aplicado diferentes técnicas de explicabilidade, como a AHN, pode ser facilmente transposta para um grafo, na qual, permite utilizar diferentes formas de interpretações dos resultados gerados, como métodos de regras, visualização do próprio grafo e árvores de decisão. A qualidade da explicação foi avaliada através da comparação com a análise da saída de outros modelos.

No estudo "*Textual eXplanations for Intuitive Machine Learning*"(**??**) desenvolve um modelo que permite ao usuário interagir textualmente com os resultados de um modelo de aprendizado de máquina, utilizando uma abordagem de perguntas e respostas (question-answering).

Para a pesquisa, foram empregados diversos conjuntos de dados, todos relacionados à área da saúde. Para a geração de textos técnicos, foi utilizado o MedDialog dataset, que contém termos médicos especializados.

A metodologia de Torri inicia com um modelo de classificação — entre eles, XGBoost, Random Forest, Regressão Logística e Redes Neurais. A partir desse modelo, aplica-se a técnica de explicabilidade SHAP para identificar as variáveis mais relevantes na classificação. As variáveis selecionadas são, em seguida, integradas a explicações gramaticais para aprimorar a precisão das interpretações geradas. Com isso, o modelo oferece explicações em formato textual, tornando os resultados mais acessíveis e intuitivos para diversos públicos. Um exemplo dessa metodologia é ilustrado na Figura 23.

3.3 Abordagens de XAI para dados de Imagens

O artigo intitulado "*Detecting mammographically occult cancer in women with dense breasts using deep convolutional neural network and Radon Cumulative Distribution Transform*", proposto por Lee e Nishikawa (2019), aborda sobre a detecção de câncer oculto em mamografias com tecido mamário denso, a abordagem se utiliza de redes neurais convolucionais profundas (CNN), em particular o modelo *VGG16* e a transformada cumulativa da distribuição de Randon.

O objetivo principal deste artigo, é melhorar a identificação de câncer, onde não pode ser visível nos métodos tradicionais, em especial, em mulheres com densidade mamária elevada. A utilização de tais técnicas combinadas, indicam uma melhor precisão na detecção do problema, complementando assim métodos já existentes.

Para as técnicas de explicabilidade, foi utilizado a técnica de Grad-CAM, que permite visualizar as áreas de maior interesse para a classificação. O Grad-CAM utilizou as áreas detectadas nas imagens para verificar se as regiões de interesse correspondem a características clínicas conhecidas, como lesões e anomalias.

O estudo de Brunese et al. (2020), intitulado "*Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays*"explora o uso de técnicas de aprendizado profundo para detectar COVID-19 em imagens de raio-X, empregando o modelo VGG16 para essa tarefa, a figura 24 exibe a implementação prática do Grad-CAM ao conjunto de dados.

Assim como no estudo de Lee e Nishikawa (2019), a técnica GRAD-CAM é utilizada para destacar as regiões de interesse nas imagens, fornecendo informações visuais que podem ser úteis para radiologistas e patologistas ao identificar áreas marcadas pelo algoritmo, como visualizado na figura 25.

O artigo *"Endoscopic Image Classification Based on Explainable Deep Learning"* Mukhtorov et al. (2023), examina a aplicação de técnicas de inteligência artificial explicável (XAI) na classificação de imagens endoscópicas. Foram empregados diferentes modelos de classificação de imagens, incluindo DenseNet201, MobileNetv2, ResNet-152, ResNet-18 e VGG16.

Para garantir a explicabilidade, foi utilizada a técnica Grad-CAM, cuja qualidade foi avaliada através da visualização dos mapas de calor nas imagens. Além disso, o artigo menciona outras técnicas, como LIME e SHAP, para aumentar ainda mais a transparência do modelo.

O estudo de (BöHLE et al., 2019) explora a aplicação de redes neurais combinadas com técnicas de explicabilidade para interpretar decisões de classificação em imagens de ressonância magnética (MRI) relacionadas à doença de Alzheimer (AD). Para esse fim, foi utilizada a técnica de propagação de relevância camada a camada (LRP), Figura 26.

Figura 24 – Explicação via Grad-CAM. Fonte: (BRUNESE et al., 2020)

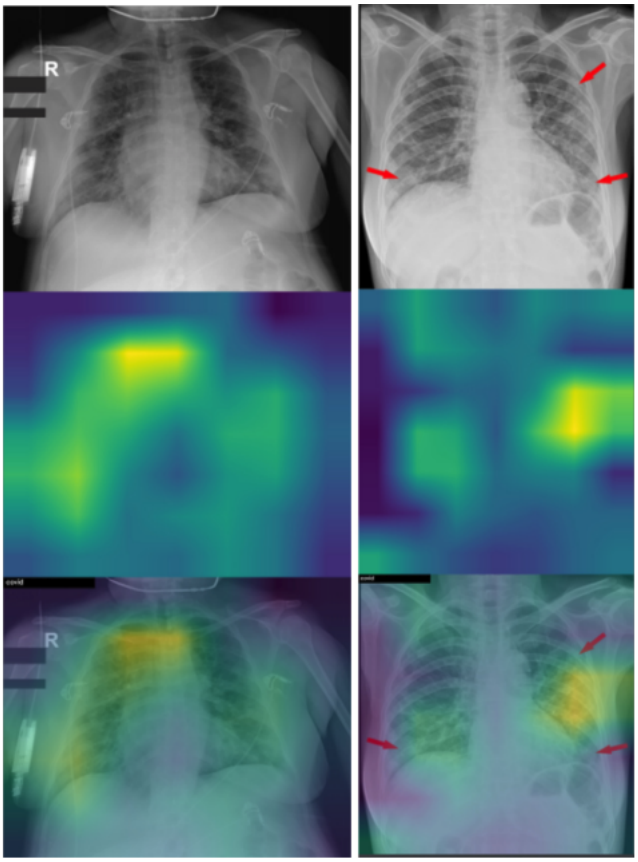


Figura 25 – Explicações via Grad-CAM. Fonte: (LEE; NISHIKAWA, 2019)

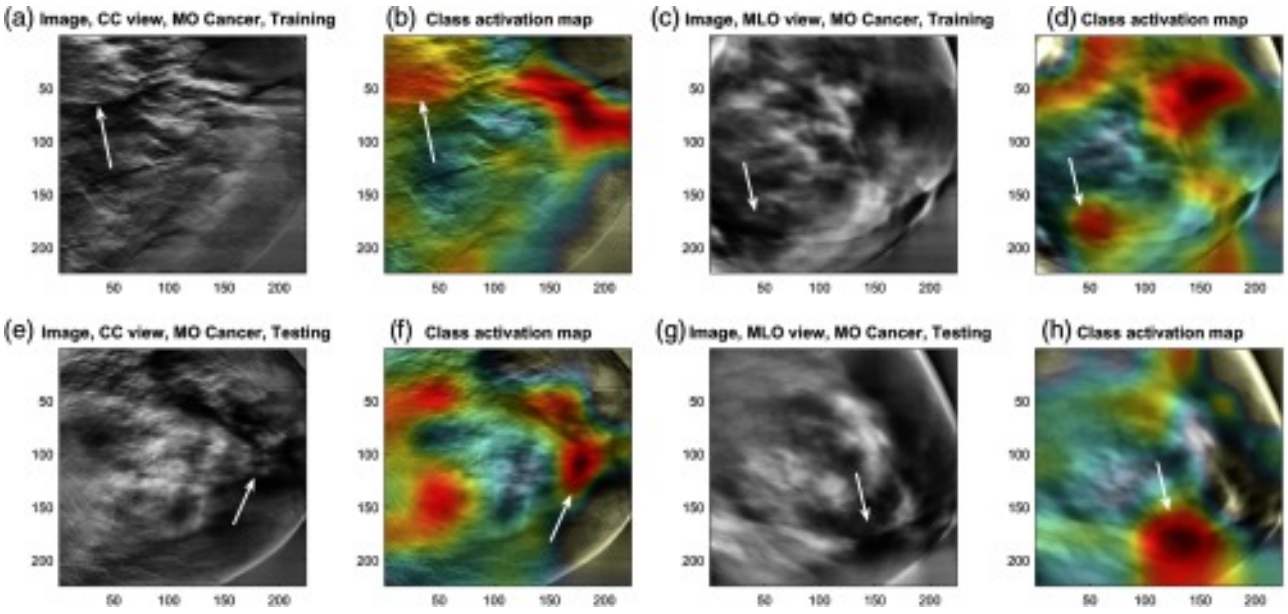
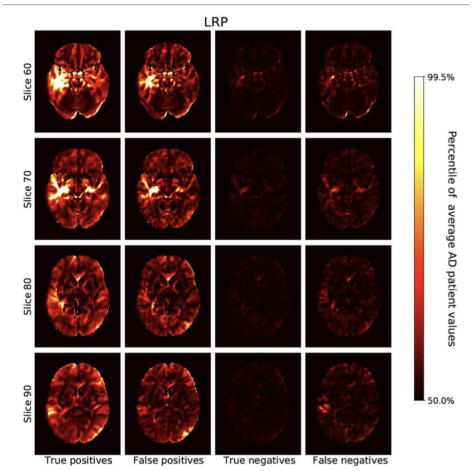


Figura 26 – Técnica de LRP. Fonte: (BöHLE et al., 2019)



Um dos principais achados deste estudo foi a significativa variabilidade entre pacientes nas áreas relevantes identificadas, indicando que a LRP pode fornecer "impressões digitais"únicas para cada indivíduo, refletindo as características específicas de sua condição.

A qualidade das explicações foi avaliada com base na sua correlação com conhecimentos prévios, destacando regiões do lobo temporal e do hipocampo, amplamente reconhecidas na literatura como associadas à patologia do Alzheimer.

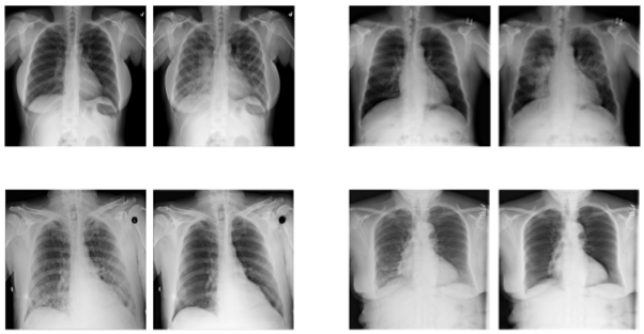
O estudo *"GANterfactual - Counterfactual Explanations for Medical Non-Experts using Generative Adversarial Learning"*(MERTES et al., 2022) propõe uma abordagem inovadora para lidar com dados de imagens médicas por meio de explicações contrafactuais. Essa técnica possibilita a comparação visual entre cenários opostos, permitindo identificar as mudanças que diferenciam um paciente saudável de um paciente doente, a figura 27, exibe a geração de explicações pelo método GANterfactual.

Neste trabalho, utiliza-se uma base de dados de COVID-19, semelhante à do estudo realizado por Brunese et al. (2020). A criação dos cenários hipotéticos é feita com dados sintéticos: a partir do conjunto de treinamento, a metodologia gera novos dados e, ao realizar a tarefa de classificação, constrói um cenário oposto ao da classificação original, introduzindo pequenas modificações que levam a uma nova classificação. Isso facilita a comparação entre diferentes cenários, proporcionando uma visualização clara das variáveis que influenciam a mudança de diagnóstico.

O artigo proposto por Nauta et al. (2023), intitulado "*Neural Prototype Tree: An Intrinsically Interpretable Method for Fine-Grained Image Recognition*"apresenta o desenvolvimento do ProtoTree, um modelo de aprendizado profundo que combina a aprendizagem de protótipos com árvores de decisão para criar uma abordagem interpretável para a classificação de imagens.

O ProtoTree foi projetado para enfrentar os desafios impostos por modelos complexos em tarefas de classificação de imagens, fornecendo explicações locais e globais de maneira

Figura 27 – Geração de Explicações pelo método GANterfactual. Fonte: (MERTES et al., 2022)



estruturada. Este método é considerado intrinsecamente interpretável, pois sua estrutura em árvore de decisão permite que as regras sejam compreensíveis e intuitivas.

No processamento de uma imagem, o modelo utiliza uma rede neural convolucional (CNN) para extrair características relevantes. A representação latente resultante é, então, direcionada por uma árvore de decisão, onde cada nó contém uma pergunta binária baseada em um protótipo, como “A imagem contém um peito vermelho?”. Dependendo da resposta, a imagem segue para o próximo nó, à esquerda ou à direita.

Além disso, o ProtoTree incorpora um sistema de roteamento hierárquico, que permite ao modelo tomar decisões de classificação de maneira organizada e interpretável, imitando o raciocínio humano. A qualidade da explicação fornecida pelo algoritmo ProtoTrees, foi avaliada pela sua capacidade de apresentar suas decisões e pela sua capacidade de entender como características visuais influenciam nas decisões.

Este processo ocorre ao receber uma imagem, a mesma é processada pela CNN que extrai as características relevantes .A representação latente gerada a partir desse processo é então utilizada para navegar pela árvore de decisão. Em cada nó, o modelo realiza uma pergunta binária baseada em um protótipo, como "A imagem possui um peito vermelho?". Conforme a resposta (sim ou não), a imagem é direcionada para o próximo nó, à esquerda ou à direita.

No artigo "*Explainability of deep neural networks for MRI analysis of brain tumors*"Zeineldin R.A. (2022). propõem o framework NeuroXAI, que integra diversas técnicas de explicabilidade para tornar redes neurais mais transparentes aos profissionais da saúde.

Focado na análise de imagens de ressonância magnética (MRI), o NeuroXAI foi desenvolvido para fornecer explicações visuais que auxiliam na interpretação das decisões de modelos de aprendizado profundo. A abordagem foi aplicada em duas tarefas principais no domínio de imagens cerebrais: classificação e segmentação.

A ferramenta oferece uma variedade de métodos de explicabilidade visual, como SmoothGrad, GradCAM, e Guided Grad-CAM — todos variantes do algoritmo CAM. Como esses métodos são aplicados após o treinamento do modelo, o *framework* oferece

uma explicação intrínseca. A utilização de múltiplos métodos de explicação permite a comparação entre as explicações geradas, aumentando a robustez e a confiabilidade na avaliação.

3.4 Análise de Trabalhos de XAI com Base em uma Taxonomia

A taxonomia de XAI proposta por Timo Speith estrutura a área de Explainable Artificial Intelligence (XAI) por meio de uma classificação sistemática que organiza as diversas metodologias de interpretabilidade. Com esse modelo, acadêmicos e profissionais podem entender melhor as abordagens existentes, escolhendo metodologias mais adequadas para diferentes contextos e necessidades.

Essa classificação é agrupada em ramos, como ilustrado na imagem, permitindo uma visão clara dos principais enfoques de XAI. Os métodos são divididos em categorias baseadas no funcionamento dos modelos (como técnicas de perturbação local para avaliar a importância de características), no tipo de resultado gerado (como saídas visuais ou textuais voltadas para o usuário), e em aspectos conceituais (como a distinção entre métodos ante-hoc e post-hoc, ou entre métodos específicos e agnósticos ao modelo).

Esses agrupamentos facilitam a escolha de métodos de XAI ao alinhar as técnicas com os objetivos específicos dos usuários. Com base nos estudos detalhados, a tabela 1,f ilustra a posição de cada elemento na taxonomia de Timo Speith, indicando como cada metodologia se encaixa dentro dessa estrutura sistematizada.

3.5 Visão dos Usuários

Meske et al. (2020) em seu trabalho "***Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities***"identificam desafios e oportunidades relacionados ao uso da explicabilidade. Um dos principais debates sobre o uso da IA é o risco desses modelos conterem vieses, que podem ser manifestados de diferentes formas.

Podemos enfrentar vieses tanto do lado humano, ao confiar cegamente nas decisões tomadas por tais modelos, conhecidos como "viéses de automação"(MESKE et al., 2020). Esse fenômeno pode levar à rejeição de possíveis erros gerados pela máquina devido ao excesso de confiança Goddard, Roudsari e Wyatt (2011).

Por outro lado, existem vieses gerados pelo próprio sistema, que podem ser intencionais ou involuntários em relação aos resultados. Visto que os dados utilizados para a criação de tais sistemas, podem conter vieses humanos que estão inerentes em suas bases

Tabela 1 – Distribuição dos Artigos com base na Taxonomia de Timo Speith (SPEITH, 2022).

Nome do Artigo:	Modelos Utilizados	Dataset	Técnicas de XAI	Tipo de Dado	Escopo	Ante-hoc/Pos hoc	Model Specific/Agnostic	Output format
Interpretable Machine Learning: A Case Study of Healthcare	Random Forest, Gradient Boosting	Sylhet Diabetes dataset, ICU Dataset	LIME, SHAP	Tabular	Local	Pos-hoc	Model Agnostic	Visual/Graphic
Clinically Applicable Machine Learning Approaches to Identify CKD	Random Forest, Gradient Boosting, XGBoost, Logistic Regression, SVM	CKD dataset, ICU	SHAP	Tabular	Global	Pos-hoc	Model Agnostic	Visual/Graphic
Machine learning-based prediction of in-hospital mortality	Logistic Regression, Random Forest, MLP, GB	University of Tokyo Hospital data	SHAP	Tabular	Global	Pos-hoc	Model Agnostic	Visual/Graphic
Generating fuzzy rules for constructing interpretable diabetes classifier	ANFIS, FCM	Pima Indian Dataset	FCM–ANFIS	Tabular	Local	Ante-hoc	Model Specific	Rules
Interpretability of Artificial Hydrocarbon Networks for Breast Cancer Classification	AHN-model	Wisconsin Breast Cancer UCI	Molecule Center Similarity	Tabular	Global	Ante-hoc	Model Specific	Rules
Detecting occult cancer in dense breasts using CNN and Radon Cumulative Distribution	VGG16	Mammogram dataset	CAM	Images	Local	Pos-hoc	Model Agnostic	Visual/Graphic
Explainable Deep Learning for COVID-19 Detection from X-rays	VGG16	COVID-19 image data, NIH Chest X-Ray	Grad-CAM	Images	Local	Pos-hoc	Model Agnostic	Visual/Graphic
Endoscopic Image Classification with Explainable Deep Learning	DenseNet201, MobileNetv2, ResNet-152, VGG16	Kvasir	Grad–CAM variants	Images	Local	Pos-hoc	Model Agnostic	Visual/Graphic
LRP for Explaining DNN in MRI-Based Alzheimer's Disease Classification	Custom CNN	Alzheimer's ADNI database	Layer-Wise Relevance Propagation (LRP)	Images	Local	Pos-hoc	Model Agnostic	Visual/Graphic
GANterfactual - Counterfactual Explanations for Non-Experts	AlexNet	RSNA Pneumonia Detection	GanterfactualImages		Local	Pos-hoc	Model Agnostic	Visual/ Graphic/ Counterfactual
Textual Explanations for Intuitive Machine Learning	XGBoost, Random Forest, Logistic Regression, FFNN	MedDialog, Cardiovascular, Pima diabetes datasets	Textual	Tabular	Global/Local	Pos-hoc	Model Agnostic	Textual
Neural Prototype Trees for Interpretable Fine-grained Recognition	Neural Prototype Tree	CUB-200-2011, Stanford Cars	ProtoTree	Images	Global/Local	Ante-hoc	Model Specific	Visual/Graphic
Explainability of DNNs for MRI Brain Tumor Analysis	ResNet-50	BraTS challenges 2019, 2021	SmoothGrad, GradCAM, Guided Grad-CAM	Images	Local	Pos-hoc	Model Agnostic	Visual/Graphic

levando a um modelo tendencioso, seja a raça, gênero, religião etc (CALISKAN; BRYSON; NARAYANAN, 2017).

Portanto, as técnicas de explicabilidade ajudam a compreender e mitigar possíveis vieses. Elas fornecem um entendimento sobre os resultados do algoritmo, de modo que diferentes usuários, com suas necessidades específicas, possam obter informações que a explicabilidade deve oferecer, facilitando assim o processo de tomada de decisão.

Por exemplo, diferentes pessoas de negócio, exigem diferentes níveis de explicação, por exemplo, usuários comuns, estão mais interessados em recursos para compreender o

raciocínio do modelo, trazendo o conhecimento gerado para o seu mundo, a fim de analisar a sua validade e confiabilidade. Enquanto gestores, estão mais propensos a necessidade de supervisionar o algoritmo e o seu uso (MESKE et al., 2020)

Também como abordado por (JUNG et al., 2023), em seu estudo para a relação de pessoas de negócios com tais modelos, é possível entender que cada membro (por exemplo, médicos, enfermeiros, pacientes) exibiam expectativas diferentes em relação a saída do modelo, isto está correlacionado com diferentes formações, treinamentos e experiências, tornando a geração de uma boa explicabilidade desafiadora.

A AI pode ter um impacto significativo na sociedade e, portanto, é importante garantir que ela seja usada para o bem comum e não prejudique as pessoas ou a sociedade como um todo, logo a construção de um sistema de IA responsável é de fundamental importância, segundo o artigo "***Towards a Roadmap on Software Engineering for Responsible AI***"(LU et al., 2022), propõem um conjunto de principais para uma IA ser responsável, nas quais são, privacidade, responsabilidade, segurança e proteção, transparência e explicabilidade, equidade e não discriminação.

Vários fatores devem ser considerados para garantir que os sistemas de IA sejam seguros, seguros, transparentes, justos e respeitosos aos valores humanos. Os indivíduos que lideram cada etapa do ciclo de vida do sistema de IA devem ser identificados e responsabilizados pelos resultados. A segurança operacional deve ser mantida ao longo do ciclo de vida e fornecer uma explicação para as decisões tomadas. Inclusão e não discriminação são essenciais para evitar tratamentos injustos. Um processo de contestação garante o controle humano, e os sistemas de IA devem defender os direitos humanos, a diversidade e o benefício coletivo (LU et al., 2022).

Para a redução de vieses no processo arquitetural, é necessário a organização de um sistema de *feedback* por parte dos usuários, além da inclusão de técnicas de *fairness*„ na qual um novo modelo seja dedicado apenas a um grupo de usuários

3.6 Conclusão

Esta revisão narrativa teve o propósito de analisar os trabalhos relacionados a esta pesquisa, que tem o objetivo de propor uma arquitetura baseada na estrutura de AIaaS, com um enfoque específico na explicabilidade e transparência dos modelos de aprendizado de máquina aplicados em cenários de diagnóstico médico. A necessidade de garantir que os modelos de saúde sejam compreensíveis e confiáveis é fundamental para sua adoção segura e eficaz em ambientes clínicos.

Na seção 3.1, foi abordado algumas arquiteturas voltadas para o domínio de aprendizado de máquina, sendo elas baseadas no paradigma KaaS, ou *pipelines* padrões de desenvolvimento de aprendizado de máquina. Apesar que nos artigos analisados, os

componentes subjacentes são uniformes, abrangendo desde a fase de pré-processamento dos dados, que engloba a coleta de informações brutas e tratamento de dados, até as etapas de treinamento do modelo, implantação.

As arquiteturas expostas são moldadas conforme a natureza do problema a ser resolvido, sendo que seus componentes têm em comum a camada de explicabilidade que visa o entendimento a partir de um modelo específico a geração de explicabilidade, garantindo uma maior transparência.

No entanto, apesar de as técnicas de explicabilidade oferecerem diferentes formas de explicação, muitas vezes a escolha dessas técnicas não segue critérios claros, como uma proposta de taxonomia ou uma definição rigorosa do conceito de explicabilidade. No contexto do presente trabalho, a arquitetura proposta busca preencher essa lacuna ao adotar a taxonomia de (SPEITH, 2022), como embasamento teórico. Essa abordagem tem como objetivo gerar explicações mais estruturadas e consistentes, alinhadas com uma fundamentação conceitual sólida, que potencialize a interpretabilidade e a utilidade das inferências realizadas.

Um papel bastante importante visto nos artigos de estudo, foi a importância do *feedback* e a participação de *stakeholders* no processo de desenvolvimento das arquiteturas e modelos. Essa participação revela-se essencial, visto que pode resultar em um desempenho aprimorado, aumentar a confiança entre a equipe médica e melhorar os resultados para os pacientes. Tal engajamento é necessário tanto na fase de especificação da arquitetura quanto na consideração das condições para re-treinamento, incluindo frequência, tratamento dos dados, *feedback* dos usuários e o ambiente operacional desejado. Ademais, o *feedback* dos usuários pode ser aproveitado para enriquecer a transparência e explicabilidade do sistema de IA, permitindo que os usuários compreendam o processo de tomada de decisões e o funcionamento do sistema.

A integração da explicabilidade e transparência não é apenas uma mera preocupação técnica, mas também está intrinsecamente ligada a princípios éticos. O acesso a informações claras sobre como um modelo toma suas decisões é vital para evitar resultados injustos ou preconceituosos, além de proporcionar um ambiente onde a responsabilidade pelo funcionamento do modelo é compartilhada entre desenvolvedores, profissionais de saúde e pacientes. Isso também fortalece a confiança na tecnologia e nos diagnósticos propostos.

No que tange à comunicação com os usuários finais, predominantemente, essa interação é intermediada através de APIs REST. Essa abordagem facilita o acesso às funcionalidades dos modelos, permitindo a troca de dados e informações de maneira eficaz e padronizada.

De maneira geral, esta revisão narrativa destaca que a proposta de uma arquitetura de AIaaS centrada na explicabilidade, transparência e princípios éticos não somente contribui para aprimorar a qualidade e confiabilidade dos modelos de diagnóstico médico,

mas também reforça a importância de uma abordagem centrada no paciente, ética e responsável para a aplicação da inteligência artificial na área da saúde.

4 H-XAIaaS (*Health - eXplainable Artificial Intelligence as a Service*)

A falta de transparência, confiança e interpretabilidade é uma das principais barreiras para a adoção de modelos ML na área médica. Para os usuários, é fundamental compreender como os modelos funcionam e como chegaram a suas conclusões. Isso é essencial para a confiança dos usuários nos modelos e para a tomada de decisões clínicas informadas.

Nos últimos tempos, houve um foco significativo no desenvolvimento de modelos de Aprendizado de Máquina direcionados ao campo da saúde. Esses modelos têm sido direcionados para resolver os mais variados problemas, tais como identificar áreas afetadas por tumores cerebrais em imagens de ressonância magnética (TU et al., 2016), e para detectar casos de Covid-19 (BRUNESE et al., 2020).

Apesar de tais modelos têm demonstrando bons desempenhos. No quesito de implementação prática desses modelos é bastante deficiente. Essa situação é influenciada por diversos fatores, incluindo questões relacionadas à padronização dos dados, compreensão limitada dos algoritmos, possíveis vieses inerentes, ameaças de segurança cibernética, e até mesmo a evolução das tendências nos dados ao longo do tempo, conceitualmente denominado de *data drift.*

Em aplicações no mundo real, os dados estão em constante mudança, inclinando a que modelos de aprendizado de máquina sejam re-treinados periodicamente, ou no pior dos cenários, que toda a *pipeline* dos modelos precise ser refeita.

No contexto da saúde, diversos tipos de dados podem ser considerados, abrangendo informações estruturadas, como registros eletrônicos de pacientes e dados não estruturados, como imagens radiográficas. A arquitetura deve tratar cada tipo de dado de forma independente, respeitando suas características próprias, para permitir uma integração ou alteração mais fluida durante o tratamento dos dados.

No entanto, tanto a arquitetura proposta por Yao et al. (2017) quanto a de Ribeiro, Grolinger e Capretz (2015) não incluem módulos de explicabilidade. Isso pode estar relacionado à necessidade recente de integrar técnicas de explicabilidade no desenvolvimento de modelos de aprendizado de máquina, uma necessidade que tem se intensificado com o crescimento dos modelos nos últimos anos. Essa demanda surge tanto da busca por maior transparência e aceitação dos modelos quanto dos requisitos regulatórios.

No decorrer deste capítulo será detalhada a arquitetura de referência para explicabilidade na área da saúde. Tal arquitetura é baseada no paradigma de inteligência artificial como serviço em conjunto com o paradigma de conhecimento como serviço (XU; ZHANG,

2005; LINS et al., 2021). A arquitetura tem como objetivo garantir maior transparência no entendimento do processo decisório do modelo, além de garantir robustez e conformidade aos princípios éticos e legais.

4.1 XH-KaaS (*eXplainable - Health Knowledge as a Service)*

O paradigma KaaS, promove a centralização do acesso ao conhecimento através de serviços para diferentes domínios, quando combinada com técnicas de explicabilidade se tornam verdadeiros aliados, pois tais técnicas oferecem um embasamento maior acerca do conhecimento extraído, aprimorando a sua aplicabilidade e confiança.

A seguir propomos a arquitetura de referência para oferecer explicabilidade como serviço no domínio da saúde. A figura 28 exibe a arquitetura conceitual proposta XH-KaaS, método proposto por (MONTENEGRO; LINO, 2024), onde observa-se a existência de dois componentes principais, o extrator de explicabilidade (do inglês, *Explanability Extractor*) e o provedor de explicabilidade (do inglês,*Explanability Provider*).

Adicionalmente, temos os Detentores de Dados (do inglês, *Data Owners*), que para o domínio médico podemos considerar prontuários eletrônicos, planilhas, imagens, diretrizes médicas, etc. Além disso, temos *insights*, o conhecimento a priori do problema que possa aprimorar os métodos de aprendizagem de máquina.

Já os Consumidores do Conhecimento (do inglês, *Knowledge Consumer*), de modo geral são usuários. Para o domínio da saúde seriam tanto profissionais da área da saúde, quanto pacientes; podendo também o conhecimento gerado ser consumido por outros serviços. Sendo possível a criação de soluções específicas e direcionadas a cada parte de interesse, como por exemplo, prover explicações direcionadas e específicas a determinados usuários.

A Figura 29 apresenta em detalhes a arquitetura XH-KaaS e seus respectivos componentes. A construção e definição desses componentes, presentes tanto no *Explanibility Extractor* quanto no *Explability Provider*, foram fundamentadas a partir da taxonomia de explicabilidade proposta por Speith (2022), a qual oferece uma visão ampla e estruturada das definições e métodos de explicabilidade.

4.1.1 Extrator de Explicabilidade (*Explainability Extractor*)

O Extrator de Explicabilidade é responsável pelo processamento e a comunicação dos dados fornecidos como entrada ao servidor provedor de explicabilidade, processo realizado por meio de consultas realizadas pelos consumidores da informação, através de inferências e/ou modelos de aprendizado de máquina.

Figura 28 – Arquitetura de Conceitual (MONTENEGRO; LINO, 2024).

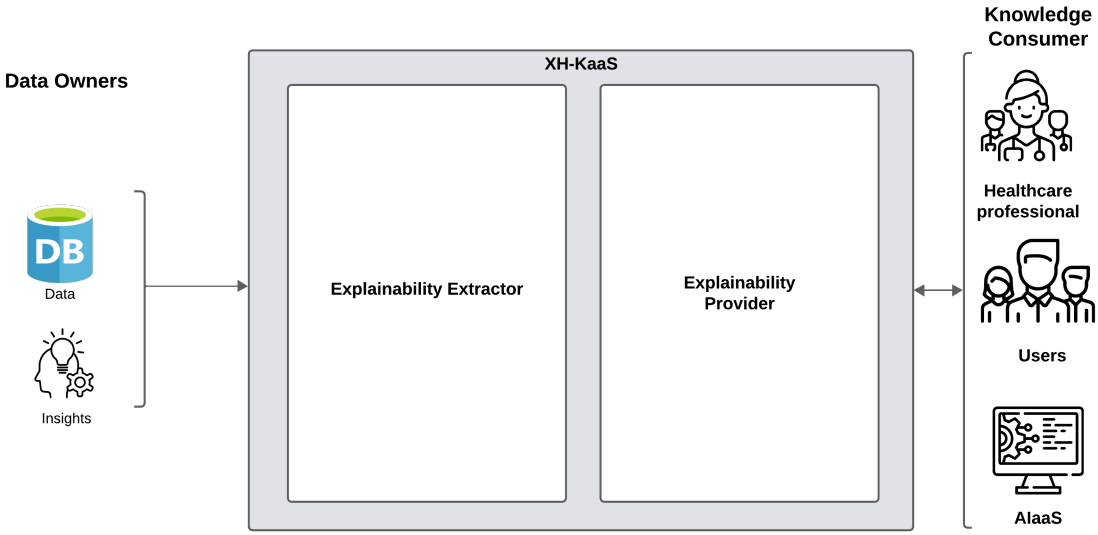
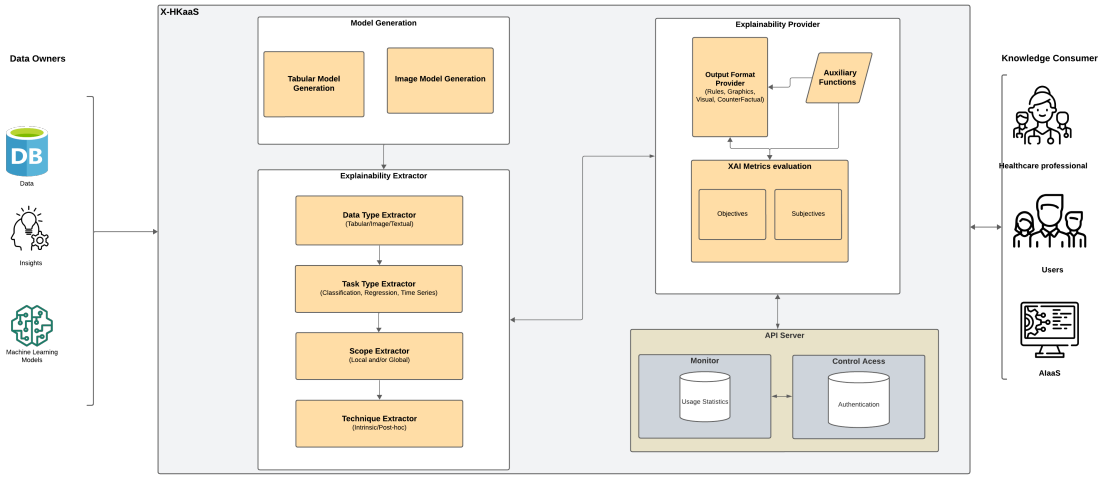


Figura 29 – Arquitetura detalhada (MONTENEGRO; LINO, 2024).



O ***Data Type Extractor*** é crucial para identificar o tipo de dado a ser explicado, a partir de quais as técnicas necessárias para garantir a explicabilidade adequada. Por exemplo, imagens médicas requerem abordagens específicas, como mapas de calor, enquanto dados tabulares exigem métodos distintos.

Este componente se relaciona com o ***Task Type Extractor*** para selecionar as melhores técnicas com base na tarefa preditiva a ser realizada. O ***Scope Extractor*** define o tipo de explicação a ser fornecida, dividindo-se em local (para uma instância) e global (para o comportamento do modelo em toda a base de treinamento).

Caso não seja fornecido nenhum modelo de aprendizado de máquina, o módulo *Model Generation*, é responsável por gerar um modelo de acordo com a tarefa a ser realizada, classificação ou regressão. Esta geração dos modelos é gerada de forma automática através das técnicas de **automl**, seguindo os critérios estabelecidos pelos usuários.

Por fim, o componente ***Technique Extractor***, método que define qual método de explicabilidade a ser usado. Pois, dependendo do modelo fornecido, será possível extrair a explicabilidade do próprio modelo, abordagem intrínseca. Caso contrário, será aplicado uma técnica *pós-hoc*, métodos que permitem extrair a explicabilidade de qualquer modelo utilizado.

4.1.2 Provedor de Explicabilidade (*Explanability Provider*)

Componente responsável por fornecer o conhecimento extraído é ***Output Format Provider***, pois de acordo com os extrator de explicabilidade e a consulta realizada pelo consumidor do conhecimento, será possível gerar a melhor explicação de acordo com a consulta realizada, o ***Output Format Provider***, oferece um conjunto de técnicas de explicação, como técnicas dedicadas a gráficos, mapas de calor, explicações contra factuais etc.

O componente responsável por fornecer o conhecimento extraído é ***Output Format Provider***, pois de acordo com os extratores de explicabilidade e a consulta realizada pelo consumidor do conhecimento, será possível gerar a melhor explicação de acordo com a consulta realizada. ***O Output Format Provide***r oferece um conjunto de técnicas de explicação, como técnicas dedicadas a gráficos, mapas de calor, explicações contra factuais, etc.

Um outro componente é o ***XAI Metrics Evaluation***. As métricas em XAI são projetadas para avaliar a qualidade, eficácia e interpretabilidade das explicações fornecidas por tais modelos (COROAMă; GROZA, 2022). Segundo (DOSHI-VELEZ; KIM, 2017), as métricas dividem a avaliação de XAI em métricas subjetivas e objetivas, como abordadas no capítulo 2, seção 2.5.

De modo geral a figura 30, exemplifica o fluxo e como os módulos apresentados interagem uns com os outros e suas respectivas ordens.

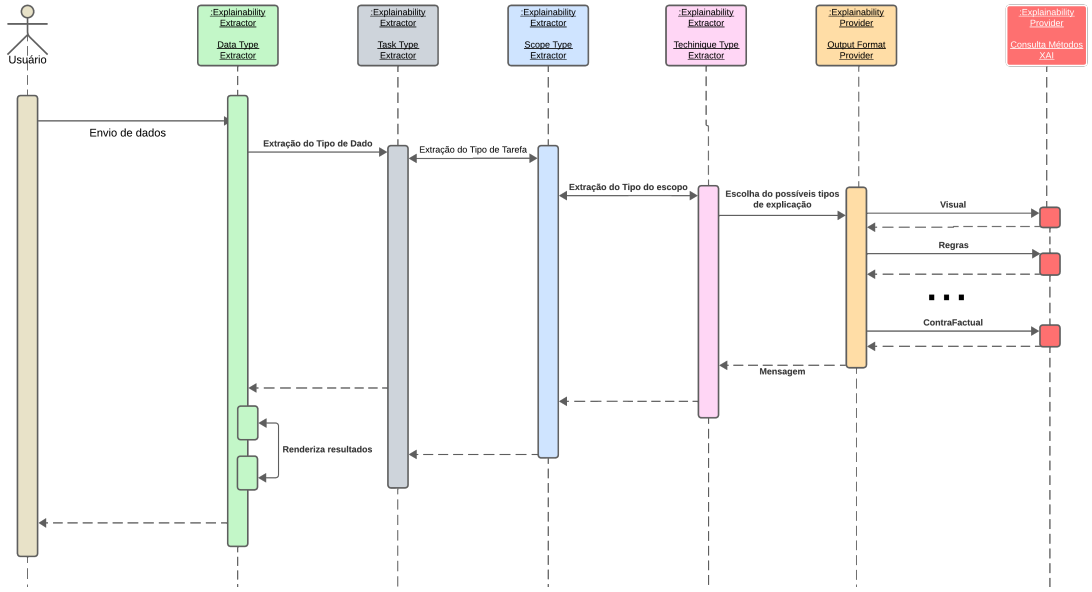
4.2 H-XAIaaS (Health - eXplainable Artificial Intelligence as a Service)

Conforme discutido na seção anterior, o paradigma **XH-KaaS** oferece uma ferramenta para o entendimento do processo decisório dos modelos, integrando o conhecimento como serviço às técnicas de explicabilidade. Para padronizar a nomenclatura, essa integração, destacada em roxo, é denominada ***XAI Provider Service***.

Com base na figura 14, que exemplifica a arquitetura **AIaaS**, apresentamos a integração com a arquitetura **XH-KaaS**, denominada **H-XAIaaS**.

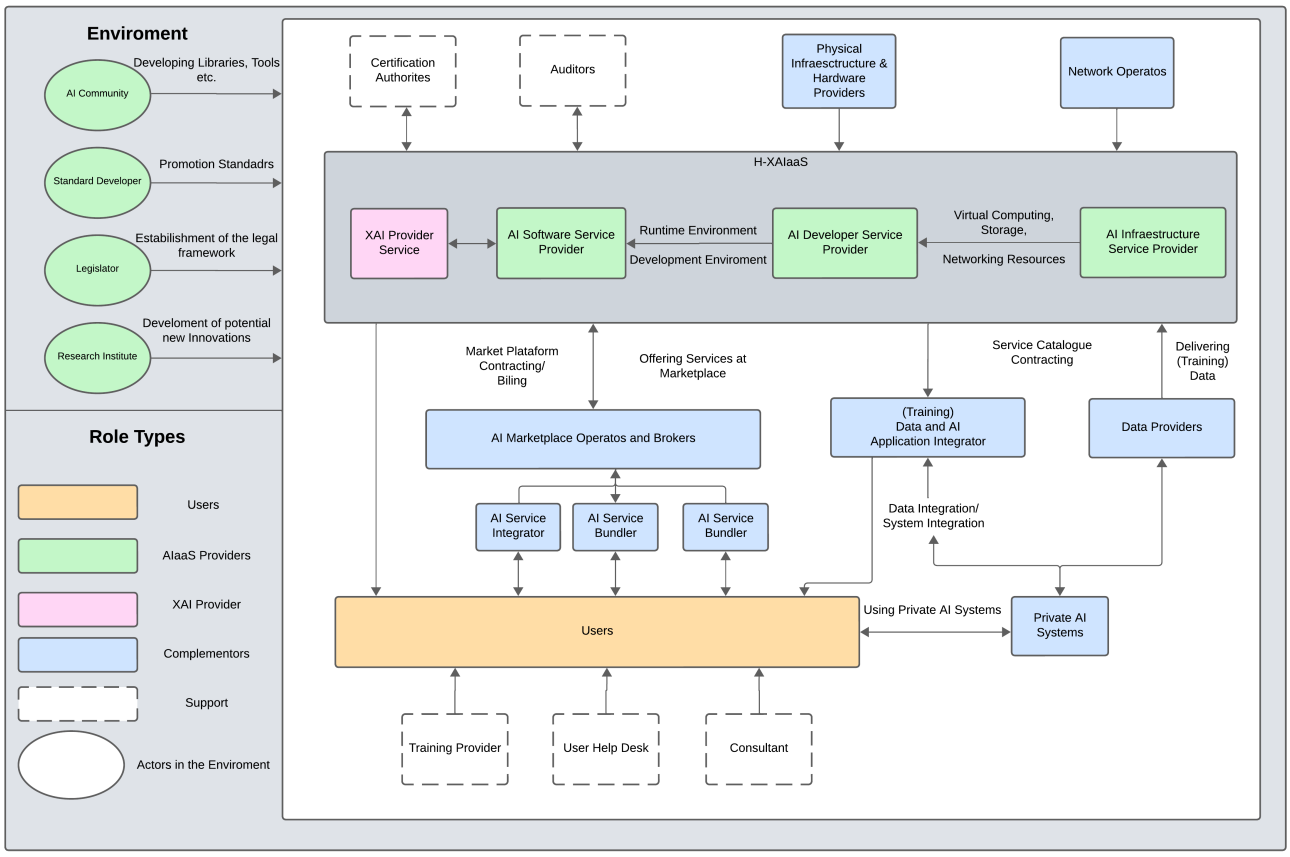
Essa integração, XH-KaaS (***XAI Provider Service***) com **AIaaS**, atua como um

Figura 30 – Diagrama Sequencial arquitetura XH-KaaS.



agente intermediário entre a requisição da saída do modelo com o usuário, proporcionando uma combinação eficiente do conhecimento técnico do usuário com a informação fornecida pelo modelo. Essa sinergia aprimora significativamente o processo de tomada de decisão, como ilustrado na Figura 31.

Figura 31 – Arquitetura H-XAIaaS. Fonte: Autoria Própria.



5 Estudo de Caso

Com objetivo de validar a arquitetura proposta, serão conduzidos dois estudos de casos, por meio de duas instâncias distintas da arquitetura proposta. Esses dois estudos têm como foco lidar com conjuntos de dados heterogêneos, compreendendo dados tabulares e dados provenientes de imagens médicas, a escolha de trabalhar com dados estruturados e não estruturados busca demonstrar a adaptabilidade da arquitetura proposta em cenários diversificados, pois cada conjunto de dados apresenta desafios e métodos diferentes para resolvê-los.

O primeiro estudo de caso empregará uma arquitetura de H-XAIaaS direcionada a conjuntos de dados tabulares. Tal abordagem se concentra respeitando o fluxo proposto na seção 4, mas direcionado a técnicas de manipulação de dados, tratamento e modelos de aprendizado de máquina relacionados a tais abordagens de dados tabulares. A arquitetura implementada será adaptada para garantir a explicabilidade e transparência dos resultados, fornecendo *insights* claros sobre as decisões tomadas a partir dos dados tabulares.

O segundo estudo utiliza uma abordagem de H-XAIaaS direcionada para manipulação de imagens médicas, também respeitando os fluxos propostos na seção 4, contudo direcionado algoritmos, técnicas de manipulação, métodos de explicabilidade para tais conjuntos de imagens.

Através desses exemplos, espera-se não apenas validar a abordagem de H-XAIaaS, mas também evidenciar como a exploração da explicabilidade e transparência pode ser aplicada de forma coerente em diferentes tipos de dados médicos. Com essa validação, busca-se contribuir para a confiança e adoção segura de modelos de aprendizado de máquina em contextos de diagnóstico médico, independente da natureza dos dados envolvidos.

Para a melhor visualização, será desenvolvido uma interface web, simulando assim a interação do usuário com os *frameworks* de explicabilidade através da nossa arquitetura proposta.

No contexto da saúde, diversos tipos de dados podem ser considerados, abrangendo informações estruturadas, como registros eletrônicos de pacientes, e dados não estruturados, como imagens radiográficas. A arquitetura deve tratar cada tipo de dado de forma independente, respeitando suas características próprias, para permitir uma integração ou alteração mais fluida durante o tratamento dos dados.

5.1 Estudo de Caso - Síndrome Respiratória Aguda Grave (SRAG)

Em 2009, foi identificado um novo tipo de vírus, do tipo influenza A H1N1. A Organização Mundial da Saúde (OMS) determinou uma emergência de saúde pública de importância internacional. No Brasil, foram confirmados cerca de 44 mil casos e cerca de 2 mil óbitos ocasionados pela H1N1.

Neste mesmo ano, o Ministério da Saúde (MS), por meio da Secretaria de Vigilância em Saúde (SVS), desenvolveu a vigilância da Síndrome Respiratória Aguda Grave (SRAG) no Brasil, com o objetivo de detectar precocemente novas epidemias ocasionadas por vírus do tipo Influenza e/ou outros vírus respiratórios.

Diante do surgimento da Sars-CoV-2 (COVID-19) em 2019, e o seu reconhecimento como agente etiológico de casos de síndrome respiratória aguda grave, a OMS declarou uma pandemia em março de 2020, que acumulau mais de 20 milhões de casos confirmados no Brasil e mais de 683 mil mortes ocasionadas por complicações da doença (BAGGIO JUSSARA A. OLIVEIRA; EXEL, 2021).

Apesar do início das campanhas de vacinação no mundo começar em 2021, apenas em 2023 a OMS decretou o fim da emergência de saúde pública de importância internacional em relação ao Sars-Cov-2, ainda sim são necessários estudos para melhor compreensão do ocorrido (MONTALVÃO, 2022). A Síndrome Respiratória Aguda Grave (SRAG) é uma condição grave que afeta os pulmões e pode levar à insuficiência respiratória, necessitando de hospitalização. A SRAG pode ser causada por diversos agentes infecciosos, incluindo vírus, bactérias e fungos.

Segundo o Ministério da Saúde, por meio de uma nota técnica (BRASIL, 2022), determinou os grupos de risco, nas quais são: "pessoas com 60 anos ou mais, população indígena, puérperas, pneumopatias, pessoas com problemas cardiovasculares, doenças hematológicas, distúrbios metabólicos, transtornos neurológicos, imunossupressão, nefropatias e hepatopatias, obesidade e pacientes com tuberculose".

Os pacientes geralmente apresentam sintomas gripais agudos graves, acompanhados de dispneia, febre alta, tosse, falta de ar, dor no peito e dificuldade respiratória. Nos casos mais graves, a SRAG pode evoluir para insuficiência respiratória, com saturação de oxigênio (O2) inferior a 95% em ar ambiente. Essa condição aumenta o risco de internação em Unidade de Terapia Intensiva (UTI) e pode exigir o uso de ventilação mecânica. Nos casos mais críticos, pode resultar em óbito (ARAUJO et al., 2020).

Por se tratar de uma doença de importância epidemiológica que deve ser notificada e investigada, surge a necessidade de compreender os principais fatores, e desenvolver um modelo de classificação que possa auxiliar na previsão de mortes por SRAG, permitindo assim a implementação de medidas preventivas mais eficazes e otimizando o atendimento

médico aos pacientes, contribuindo significativamente para a redução da mortalidade associada à SRAG.

Para o desenvolvimento de tal estudo, foram utilizados os dados clínicos de pacientes com SRAG atendidos pelo Sistema Único de Saúde (SUS)[1]. A base de dados contém registros do período de 2021 a 2024, sendo atualizada semanalmente.

5.1.1 Base de Dados

Para o presente estudo, foram utilizados os dados do período de 2021 e 2022 para o conjunto de treinamento do modelo de aprendizagem de máquina e os dados de 2023 e 2024 para validação. A base contém cerca de 166 variáveis, contendo dados demográficos, clínicos, laboratoriais, dentre outros. O dicionário completo pode ser encontrado no site oficial do OpenDataSUS[2].

Com base nos dados de SRAG, realizaremos uma tarefa de classificação para prever a evolução dos casos de pacientes diagnosticados com SRAG. A evolução dos casos é definida como a cura ou o óbito dos pacientes diagnosticados.

Na Tabela 2, pode ser encontrada a descrição de algumas variáveis encontradas no modelo.

Tabela 2 – Exemplo de variáveis presentes na base

Variável	**Descrição**
Sinais e Sintomas/Febre	1 - Sim; 2 - Não; 3 - Ignorado
Sinais e Sintomas/Saturação O2< 95%	1- Sim 2 - Não; 3 - Ignorado
Uso de suporte ventilatório?	1-Sim, invasivo 2-Sim, não invasivo 3-Não 9-Ignorado
Evolução do Caso	1-Cura 2-Óbito 3- Óbito por outras causas 9-Ignorado

A variável alvo do nosso modelo, será a classificação da evolução do caso. Por questões de simplificação do modelo e buscando obter uma melhor performance, a evolução de "Óbito por outras causas" foi agrupado em apenas "Óbito" e a classe "Ignorado", foi excluída da base, tais detalhes serão detalhados nas seções seguintes.

5.1.2 Construção do Modelo SRAG

5.1.2.1 *Pipeline* para construção do Modelo

Para a construção do nosso modelo, foi realizado as seguintes etapas abaixo:

A figura 32, exemplifica as etapas necessárias para a construção do modelo, cada etapa definida na imagem será explicada nos tópicos a seguir.

[1] SRAG 2021 a 2024 - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19: <https://opendatasus.saude.gov.br/dataset/srag-2021-a-2024>

[2] Dicionário de Dados SRAG. <https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SRAG/pdfs/Dicionario_de_Dados_SRAG_Hospitalizado_19.09.2022.pdf>

Figura 32 – Pipeline para a construção do modelo. Fonte: Autoria Própria.

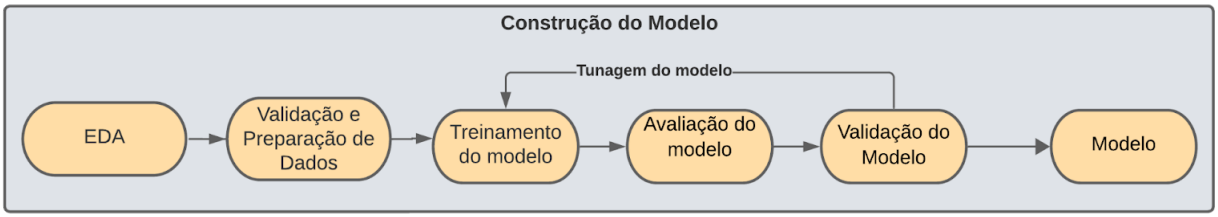


Tabela 3 – Base de pacientes filtrada por Idade

	Base Completa	Base Filtrada Idade	Diferença	Diferença Percentual
Tamanho da Base	2288766 pacientes	2286005	2761	0.12%

5.1.2.2 Análise Exploratória dos dados

Análise Exploratória de Dados (Exploratory Data Analysis - EDA), consiste em resumir e organizar os dados coletados, a partir de tabelas, gráficos, visualizações ou medidas numéricas, de forma a procurar algum padrão ou comportamento sobre as variáveis em estudo (BORNIA C., 2010). Essas visualizações permitem um entendimento mais aprofundado dos dados e ajudam a mitigar potenciais problemas na performance do modelo final, como a presença de *outliers*.

5.1.2.3 Validação e Preparação dos dados

A etapa de validação e preparação de dados, consiste na finalidade de eliminar ruídos relacionados aos dados, adequando-os para um formato desejado.

Desta forma foram realizadas algumas manipulações específicas em colunas, que serão detalhadas a seguir.

A Figura 33 ilustra as etapas do tratamento de dados. Inicialmente, a base de dados continha aproximadamente 166 variáveis. No entanto, muitas dessas variáveis apresentavam um elevado percentual de valores ausentes, o que poderia resultar em um preenchimento significativo de dados faltantes e, consequentemente, aumentar o ruído na base de dados.

Além disso, foram selecionadas variáveis com base nas diretrizes médicas que determinam os sintomas mais relevantes, conforme estabelecido pelo Ministério da Saúde (BRASIL, 2022), resultando em um conjunto final de aproximadamente 38 variáveis, listadas no apêndice A. Na ausência de informação, as variáveis categóricas foram preenchidas com a categoria "Ignorado".

Adicionalmente, foi aplicado um filtro de idade, incluindo apenas pacientes com idades entre 0 e 100 anos. Esta filtragem resultou em uma perda de menos de 1% dos dados, conforme demonstrado na Tabela 3.

Figura 33 – Etapas de Tratamento de Dados. Fonte: Autoria Própria.

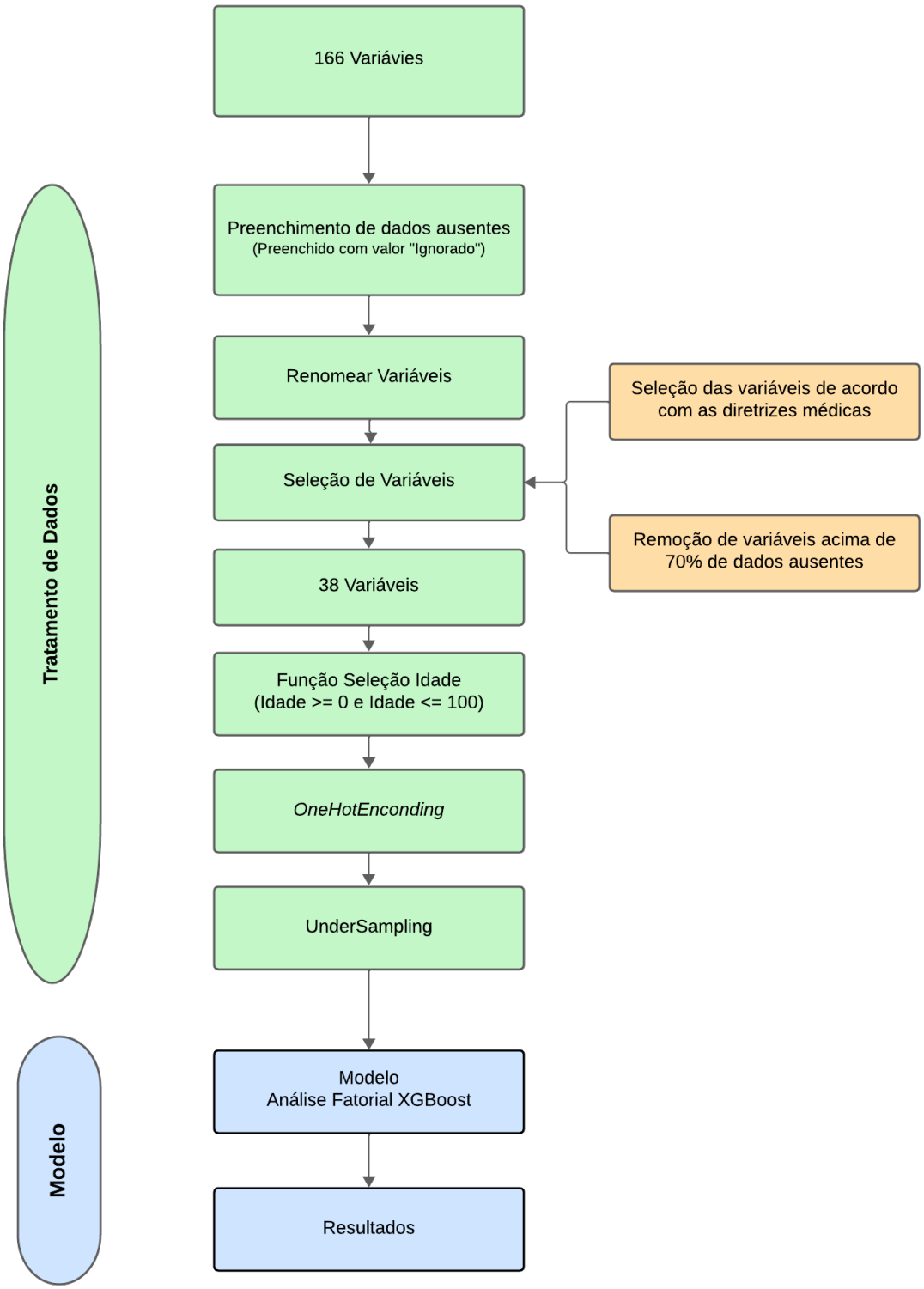


Tabela 4 – Exemplo da variável Febre sem usar a técnica One-Hot Encoding.

Sinais e Sintomas/Febre
Sim
Não
Ignorado

Tabela 5 – Exemplo da variável Febre pós aplicação da técnica One-Hot Enconding.

Sinais e Sintomas/Febre_Sim	Sinais e Sintomas/Febre_Não	Sinais e Sintomas/Febre_Ignorado
1	0	0
0	1	0
0	0	1

5.1.2.4 One-Hot Enconder

Variáveis categóricas são variáveis que os valores são definidos a partir de um grupo de categorias ou rótulos. Por exemplo, na nossa base de dados a variável gênero é dividida em três valores: masculino, feminino e indefinido. Em algumas variáveis essas categorias possuem uma ordem intrínseca, por exemplo, a escolaridade, onde tem-se: ensino fundamental, ensino superior, pós-graduação, etc.

A técnica *One-Hot Encode* (OHE), realiza a codificação de uma variável categoria em um vetor numérico. Transformando cada variável em um vetor binário de tamanho igual ao número de classe pertencentes a variável categórica, onde cada elemento no vetor é 0, exceto ao seu correspondente na categoria, marcada com 1. As Tabelas 4 e 5 exemplificam o funcionamento da técnica.

A técnica de OHE apresenta alguns benefícios, como a melhora na precisão dos modelos, pois permite que modelos trabalhem com dados categóricos de forma mais eficiente, além de melhorar a visualização. Contudo, com o aumento da dimensionalidade, a técnica de OHE aumenta o número de coluna de dados, o que pode tornar o problema mais complexo, aumentando o custo computacional, aumentando o risco de *overfitting*, como abordado na fundamentação teórica, e deteriorando o seu desempenho.

5.1.2.5 *Undersampling* (Subamostragem)

Um dos desafios ao lidar com tarefas de classificação em aprendizado de máquina é o desequilíbrio de dados. Nesse cenário, uma classe específica, chamada de classe minoritária, é representada por um número menor de instâncias em comparação com outra classe, a classe majoritária, que possui um número maior de observações (LEMNARU; POTOLEA, 2012; KOZIARSKI, 2020).

Em problemas de classificação, é comum tratar a classe majoritária como algo "negativo"ou problemático, uma vez que o foco está na classe minoritária. Esta classe

representa um evento raro e é de maior interesse ou importância em relação à classe majoritária (LEMNARU; POTOLEA, 2012).

Uma das abordagens que visam lidar com o problema de desequilíbrio de dados, ou desbalanceamento entre as classes, é a técnica de subamostragem, que consiste em equilibrar a distribuição de classes no conjunto de dados, a partir da remoção de dados pertencentes à classe majoritária de forma que iguale o número de exemplos da classe minoritária (LóPEZ et al., 2013).

A abordagem mais comum durante a aplicação do subamostragem, é a remoção aleatória dos dados da classe majoritária até que ambas as classes tenham o mesmo número de exemplos, equilibrando assim as classes. Tal equilíbrio pode ajudar a reduzir o viés dos algoritmos de aprendizado de máquina em relação à classe majoritária, permitindo a identificação de exemplos da classe minoritária e aprimorando a explicabilidade.

Contudo tal abordagem apresenta alguns riscos, visto que temos a remoção de dados pertencentes a uma classe, logo temos uma perda da informação presente na base de dados, o que impacta na capacidade de generalização do modelo diante do problema.

A figura 34, exemplificam a aplicação da técnica de subamostragem, onde temos uma classe minoritária inicialmente com cerca de 27%, e após a aplicação temos um equilíbrio entre a classe anteriormente majoritária com a classe minoritária.

5.1.2.6 Modelos Utilizados

Para a resolução do problema, foram utilizados quatro modelos distintos: *Random Forest* (Florestas Aleatórias), *Decision Trees* (Árvores de Decisão), *Extra Trees* (Árvores Extras) e XGBoost (*eXtreme Gradient Boosting*). A definição de cada um desses métodos está descrita na seção 2.

A escolha desses modelos se baseia em suas características e desempenhos conhecidos em problemas de classificação. O *Random Forest* e o *Extra Trees* são métodos baseados em ensemble que combinam múltiplas árvores de decisão para melhorar a precisão e reduzir a variância. As *Decision Trees* fornecem uma

interpretação direta das decisões baseadas em regras, enquanto o *XGBoost* é um algoritmo avançado de *boosting* que tem se destacado pela sua capacidade de lidar com dados desbalanceados e por sua eficiência em competições de *machine learning*. Cada um desses métodos contribui com diferentes abordagens e vantagens para a resolução do problema em questão.

5.1.2.7 Treinamento do Modelo

Após o tratamento dos dados, foram selecionados quatro modelos de aprendizado de máquina, conforme descrito na seção 5.1.3.5. A seleção visa identificar o algoritmo com melhor desempenho, com base nas métricas de avaliação detalhadas na seção 5.1.3.4.

Figura 34 – Comparativo aplicação técnica *Undersampling*. Fonte: Autoria Própria.

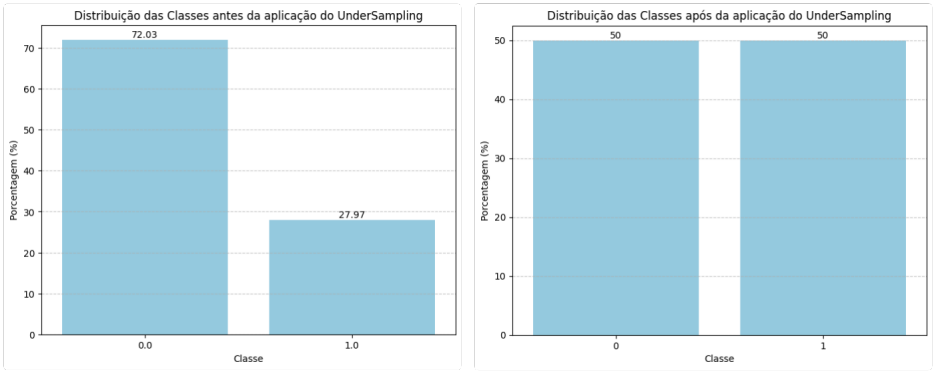


Tabela 6 – Métricas do Modelo XGBoost. Fonte: Autoria Própria.

	Precisão	Revogação	F1-Score
Classe			
0	0.77	0.77	0.77
1	0.77	0.77	0.77
acurácia			0.77

Para validar essa escolha, foi realizado um *benchmark* utilizando validação cruzada para determinar qual modelo melhor se adapta aos dados e apresenta a melhor performance.

A figura 35, exemplifica a curva ROC e os diferentes modelos empregados. Podemos observar que o algoritmo XGBoost, foi o modelo que teve melhor desempenho diante dos nossos dados.

Após a escolha do modelo, realizamos o treinamento do mesmo, utilizando a técnica de divisão de dados, dividindo em proporções de 80% e 20%. Onde 80% dos dados estava presente no conjunto de treino e 20% no conjunto de teste. Ressaltado que iremos utilizar os dados dos anos de 2023 e 2024 para treinamento e validação do modelo.

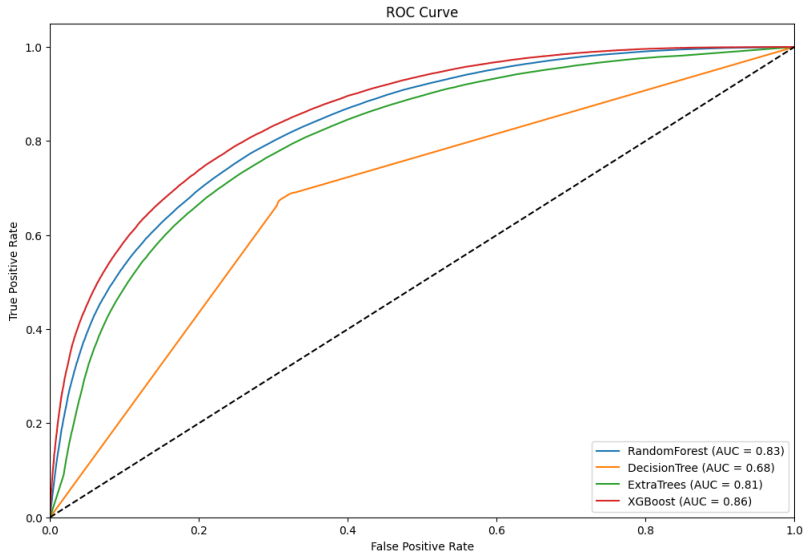
O modelo XGBoost, apresentou o seguinte comportamento para as classes, conforme apresentado na Tabela 6. De maneira geral, o modelo apresentou uma precisão de 0.77 para as três classes e 0.86 para a curva ROC.

5.1.2.8 Técnicas de Explicabilidade

Neste trabalho, foram escolhidas três técnicas distintas de explicabilidade: LIME (Local Interpretable Model-Agnostic), Anchor Rules e explicações contrafactuais. Essas técnicas foram descritas na Seção 2.4. A escolha dessas abordagens foi baseada em suas características complementares, nas vantagens específicas que cada uma oferece, e no tipo de explicabilidade que proporcionam. Além disso, a tais técnicas buscam abranger diferentes formas de visualização da explicabilidade.

A técnica LIME cria modelos interpretáveis localmente para explicar previsões

Figura 35 – Comparativo da performance de diferentes modelos na curva ROC. Fonte: Autoria Própria.



de modelos complexos, permitindo uma análise detalhada das decisões do modelo em torno de pontos de interesse individuais. Anchor Rules, por sua vez, fornece explicações baseadas em regras que permanecem verdadeiras para a maioria das instâncias semelhantes, facilitando a interpretação dos resultados e a identificação de padrões importantes.

Já as explicações contrafactuais ajudam a entender como pequenas mudanças nos dados de entrada poderiam ter levado a resultados diferentes, explorando as fronteiras de decisão do modelo e destacando quais características têm maior influência na tomada de decisão. A combinação dessas técnicas oferece uma visão abrangente das explicações, cobrindo diferentes aspectos da interpretabilidade e proporcionando uma compreensão mais completa do comportamento do modelo.

5.1.3 Aplicação e Avaliação dos Resultados

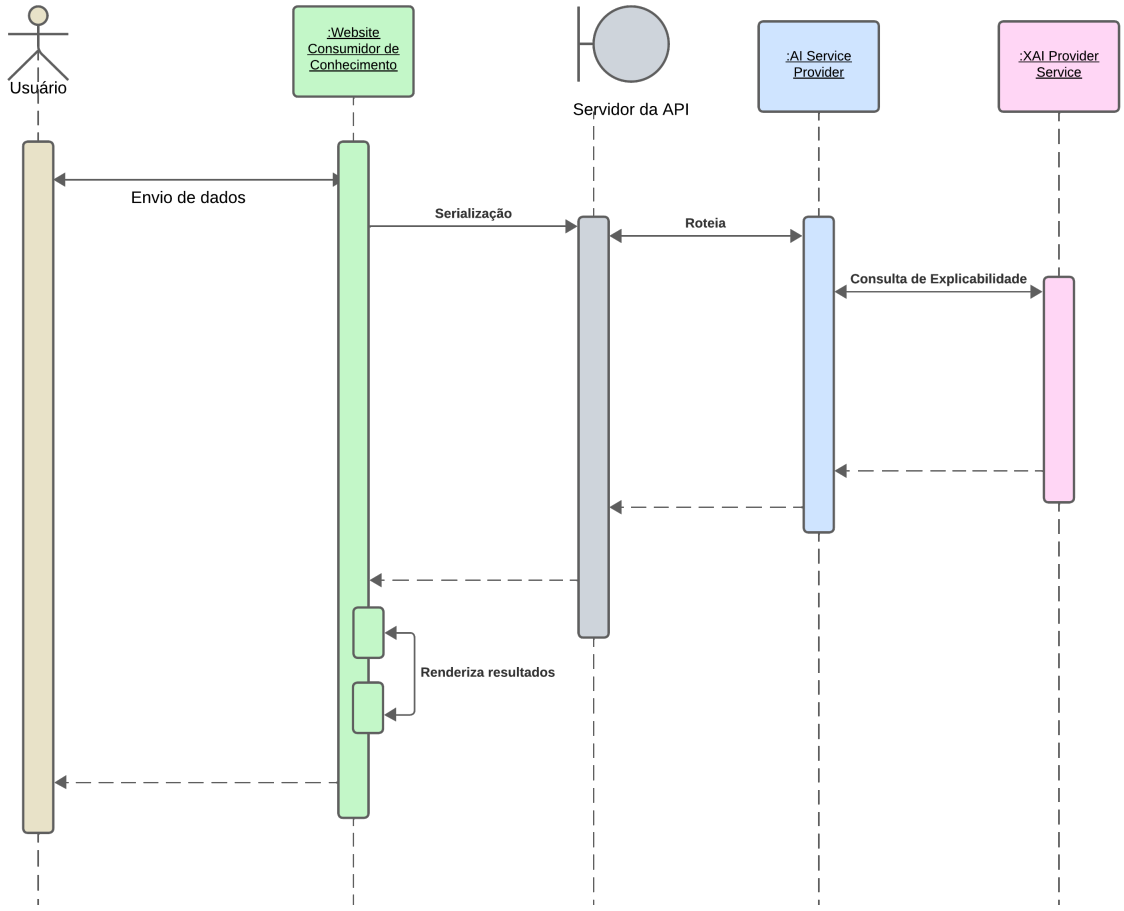
Para validação da arquitetura de referência para explicabilidade proposta neste trabalho, uma aplicação web foi desenvolvida com o objetivo de ilustrar, de maneira interativa, o funcionamento de nossa arquitetura em um sistema em produção. Esta aplicação web foi implementada em Python[3], utilizando a biblioteca Streamlit[4].

A Figura 36 apresenta o diagrama sequencial, detalhando como os diferentes objetos interagem entre si e as respectivas interações entre eles. Este diagrama ilustra o fluxo de comunicação entre os componentes os componentes da arquitetura, destacando a ordem e o método das interações.

Com base no diagrama sequencial, foi criado uma aplicação web, que busca simular as interações de um usuário com um modelo de AIaaS, como abordado na figura 4.2.1. O

[3] Pyhon. <https://www.python.org/>
[4] Streamlit. <https://streamlit.io/>

Figura 36 – Diagrama sequencial H-XAIaaS. Fonte: Autoria Própria.



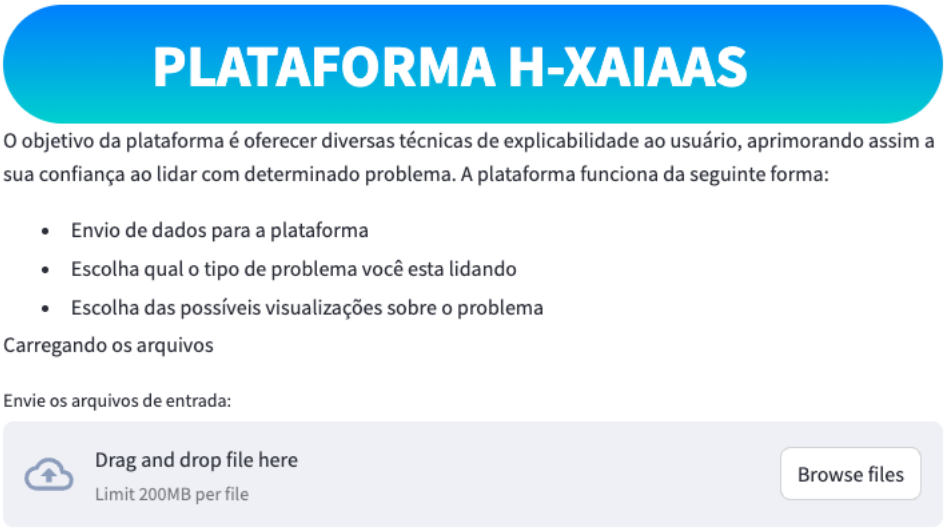
usuário interage com arquitetura, através do envio de dados, os dados são processados pela A*I Software Service Provider*, e com base no dado enviado e a técnica de explicabilidade escolhida é retornado ao usuário a explicação em conjunto com a tarefa de classificação.

A Figura 37 apresenta a página inicial da aplicação, conforme exemplificado na Figura 29. Nesta interface, o usuário interage com a plataforma por meio do envio de dados, que pode ser realizado tanto através do *upload* de diferentes arquivos (.csv, .excel, .txt, .png) quanto pela leitura de um banco de dados. Para fins de exemplificação, utilizaremos um arquivo .csv, devido a sua facilidade de envio de dados à plataforma.

Após o envio dos dados para a plataforma, eles são serializados para permitir sua interpretação, tanto na geração de predições pelo modelo disponibilizado quanto na aplicação das técnicas de explicabilidade. A partir do tipo de arquivo enviado, já no módulo *XAI Provider Service*, é possível identificar o formato do dado a ser processado, seja ele tabular ou uma imagem, e determinar as técnicas mais adequadas para esse tipo de dado, utilizando o módulo ***Data Type Extractor*** desenvolvido neste trabalho.

Após o envio dos dados, o usuário especifica o tipo de problema a ser resolvido, seja de classificação ou de regressão, utilizando o módulo ***Task Type Extractor*** da arquitetura proposta, conforme ilustrado na figura 30.

Figura 37 – Home Page da nossa aplicação. Fonte: Autoria Própria.



A Figura 39 representa o módulo ***Scope Type Extractor***, que identifica, qual instância o usuário deseja visualizar, se desejamos obter a explicação para uma instância única, ou seja, explicações locais, ou para o modelo como um todo.

Após a escolha da forma como desejamos entender a explicação do nosso problema, a arquitetura H-XAIaaS se comporta como uma sequência de filtros que, com base nas escolhas dos usuários, selecionam um conjunto de possíveis técnicas de explicabilidade para o usuário

Com base nas escolhas do usuário, é possível determinar quais técnicas de explicabilidade podem ser aplicadas. Para nosso critério de seleção, utilizamos três métodos principais (***Output Format Provider***):

- Explicações Contrafactuais;
- Explicações Visuais;
- Regras.

Dependendo da técnica de visualização escolhida, diferentes abordagens de explicabilidade podem ser identificadas e aplicadas. A Figura 40 ilustra as possíveis técnicas de explicações visuais que podem ser selecionadas.

Com base nas possíveis técnicas de visualização, vamos explorar como cada abordagem se adapta ao nosso conjunto de dados, destacando as diferenças entre os tipos de explicabilidade oferecidos por cada técnica. Isso permitirá uma compreensão mais profunda de como cada método pode revelar aspectos distintos do modelo, bem como a forma como essas técnicas podem ser aplicadas para fornecer *insights* específicos sobre o comportamento do modelo e sua tomada de decisões.

Figura 38 – Escolha do tipo da Tarefa. Fonte: Autoria Própria.

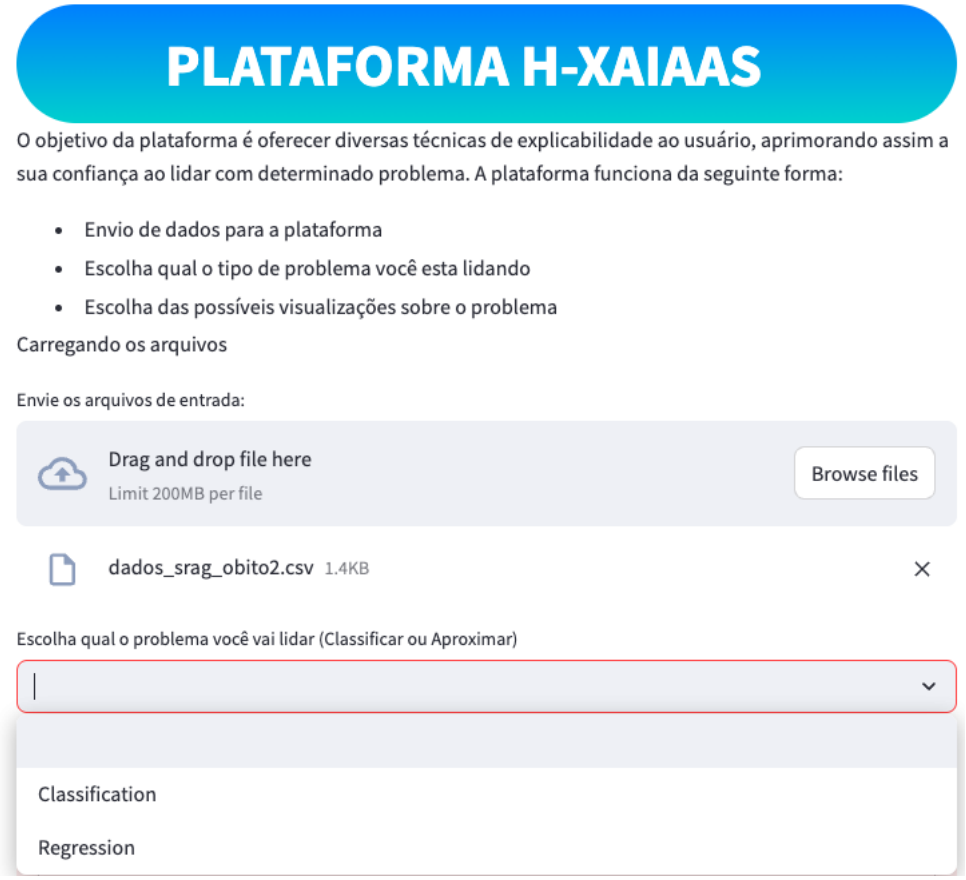


Figura 39 – Escolha do Escopo. Fonte: Autoria Própria.

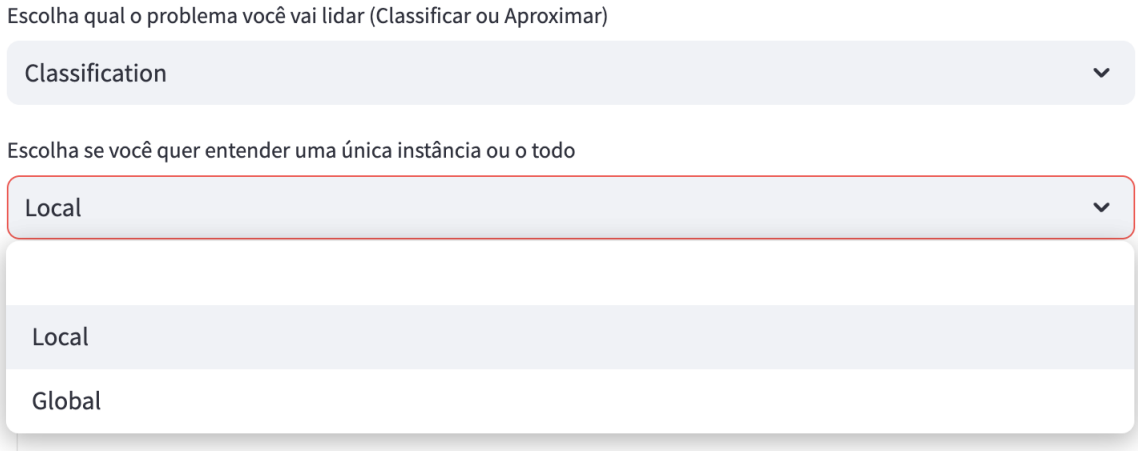
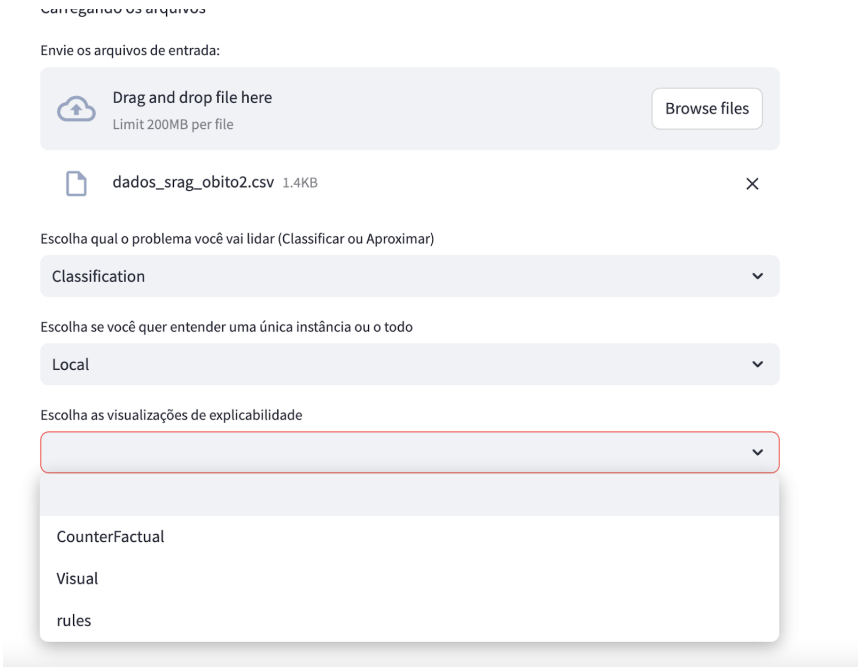


Figura 40 – Possíveis técnicas para explicações visuais a serem selecionadas. Fonte: Autoria Própria.



5.1.4 Explicação Visual

As técnicas de explicação visual, visam prover a explicabilidade através de gráficos, a Figura 41, exemplifica as possíveis técnicas de explicabilidade visual, com base nas escolhas realizadas pelo usuário, explicações locais, dados tabulares etc.

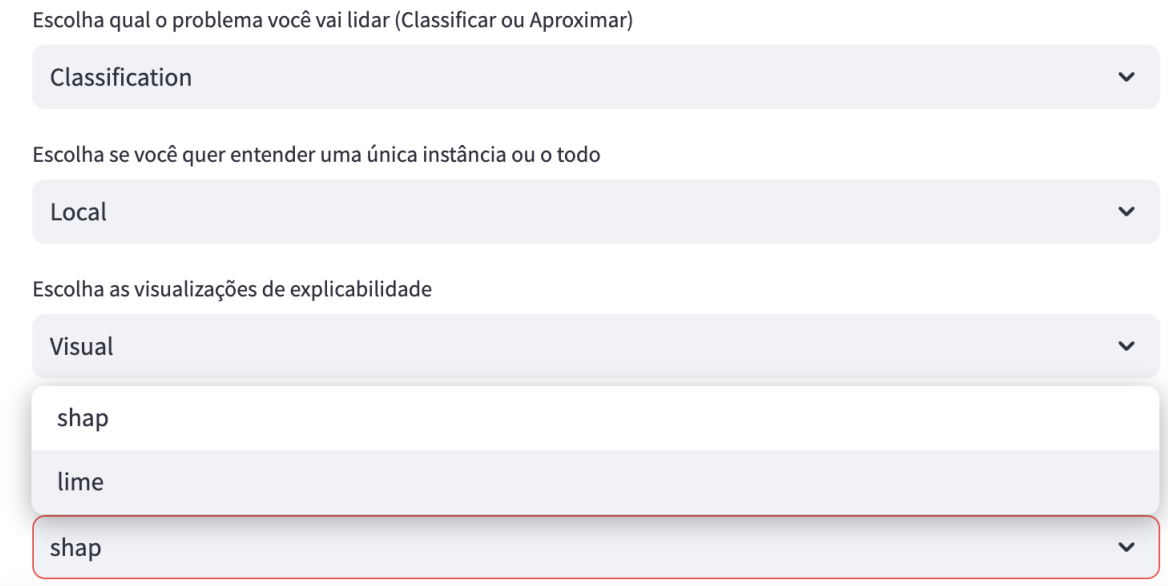


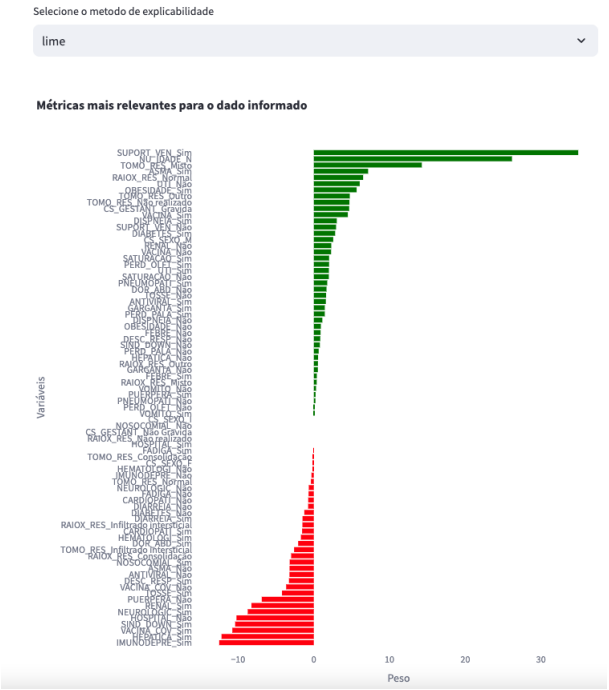
Figura 41 – Possíveis técnicas para explicações visuais a serem selecionadas. Fonte: Autoria Própria.

Ao escolher a técnica LIME, as informações baseadas nas escolhas do usuário são enviadas ao servidor. O servidor processa essas informações e classifica o dado a ser

previsto utilizando o modelo XGBoost. Com base nos dados e nas escolhas do usuário, é realizada a consulta da explicabilidade, conforme ilustrado na Figura 41.

Os resultados são então renderizados pela plataforma, que exibe a explicabilidade em formato gráfico na interface do usuário, conforme mostrado na figura 42.

Figura 42 – Explicação gerada pelo LIME. Fonte: Autoria Própria.



A Figura 42 acima mostra o gráfico de explicabilidade do modelo LIME, onde as cores verde e vermelho representam associações positivas e negativas, respectivamente para a classificação. Por exemplo, as variáveis **SUPORT_VEN_SIM** e **NU_IDADE_N** possuem uma forte carga positiva na previsão do dado, neste caso específico, de óbito. Da mesma forma, variáveis que indicam a presença de obesidade e diabetes também contribuem para esse cenário.

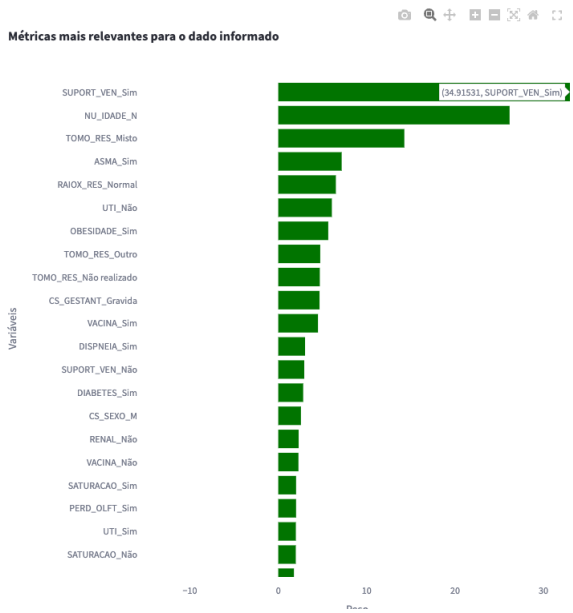
Ao observar as variáveis de maior impacto, é possível observar a **SUPORT_VEN_SIM**, com um peso de 34.55%, que indica que se o paciente utilizou suporte ventilatório, apresenta um peso de cerca de 34% para predição, como visto na figura 43.

5.1.5 Explicação via Regras

Caso a visualização do LIME não atender às expectativas, a arquitetura permite ao usuário escolher a técnica que melhor se adapta às suas necessidades. Nesse contexto, uma das possibilidades, é optar pelas regras de âncora.

A explicação gerada pelo *Anchor Rules* Figura 44, é uma explicação de regras, de modo geral, podemos interpreta a explicabilidade gerada como:

Figura 43 – Explicação gerada pelo LIME, variáveis com efeito positivo. Fonte: Autoria Própria.

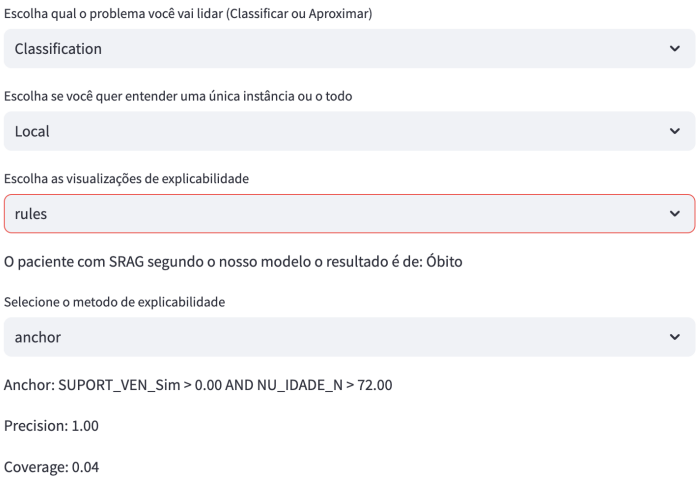


"Se o paciente utilizou suporte ventilatório e tem mais de 72 anos, então podemos confirmar a classe Óbito."

É importante destacar que essa regra apresenta uma precisão de 1.0 (100%) e uma cobertura de 0.04 (4%). Isso significa que a regra cobre 4% da base de dados e, dentro desses 4%, se o paciente atender a ambos os critérios (idade superior a 72 anos e uso de suporte ventilatório), a previsão de óbito será correta em 100%.

Vale destacar que tal definição de assertividade gerada pela predição tem influência das etapas de pré-processamento dos dados.

Figura 44 – Explicação gerada por *Anchor Rules*. Fonte: Autoria Própria.



5.1.6 Explicação Contrafactual

A Figura 45, exibe a escolha do método contrafactual. Ao escolher a explicação contrafactual, estamos em busca de um cenário oposto ao nosso problema, que neste caso nosso problema é a classificação de óbito. Estamos então buscando identificar qual seria o comportamento do modelo para o cenário da cura.

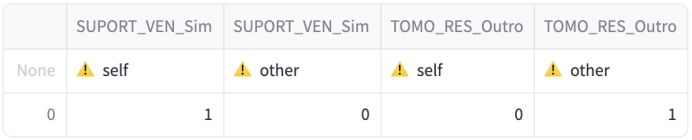
Figura 45 – Seleção do método Contrafactual. Fonte: Autoria Própria.

Uma tabela foi gerada para mostrar os dados informados, sinalizados como "***self***", em comparação com a base de dados para o cenário de cura (identificados como "***other***"). Na Figura 46, podemos visualizar este comparativo, onde fica evidente que a variável "Suporte Ventilatório"exerce um papel significativo no caso de óbito. No cenário hipotético em que o paciente não tivesse utilizado suporte ventilatório, as probabilidades de cura seriam maiores.

5.1.6.1 Avaliação da Explicabilidade

Foi possível observar, durante as três técnicas distintas de explicabilidade, uma coesão, em relação a variável de suporte ventilatório, como sempre sendo uma variável de destaque, mas como podemos saber se isso de fato compreende o que está na literatura?

Figura 46 – Explicação gerada pelo método Contrafactual. Fonte: Autoria Própria.



	SUPORT_VEN_Sim	SUPORT_VEN_Sim	TOMO_RES_Outro	TOMO_RES_Outro
None	⚠ self	⚠ other	⚠ self	⚠ other
0	1	0	0	1

Tal fato é possível ser comparado com os materiais presentes na literatura, para este projeto iremos correlacionar os *outputs* das explicações com a literatura, para projetos reais é importante a participação do ***stakeholder*** e dos especialistas de domínio, como abordado por (LANGER et al., 2021).

O estudo proposto por RANZANI (2021), denominado de "*Characterisation of the first 250.000 hospital admissions for COVID-19 in Brazil: a retrospective analysis of nationwide data*", resultou em que cerca de 80% dos pacientes que foram intubados no Brasil entre fevereiro à agosto foram a óbito. De modo similar aos estudos propostos por Huang, Lim e Pranata (2020) e Wang et al. (2024) que casos graves possuem uma maior incidência de serem hospitalizados (29%) além de uma maior mortalidade (15%).

5.2 Estudo de Caso - Classificação Osteoartrite no Joelho

Osteoratite ou artrose de joelho, se enquadra na classe de doenças agrupadas no que referimos de 'reumatismos' (REUMATOLOGIA, 2017), em um estudo proposto de Cross et al. (2014), é observado que a osteoartrite é 11º lugar de incapacidade, no Brasil é responsável por cerca de 7.5% dos afastamentos de trabalho, além de 10,5% em relação a auxílio doença no sistema previdenciário e a quarta a determinar aposentadoria (REUMATOLOGIA, 2017), visto ser a principal causa de incapacidade entre idosos (CAMANHO; IMAMURA; ARENDT-NIELSEN, 2011).

A artrose no joelho é a quarta causa mais comum de problemas de saúde entre as mulheres (RECOMMENDATIONS. . . , 2000), a mesma que se torna mais frequente após os 60 anos.

A osteoartrose do joelho é uma doença de caráter inflamatório e degenerativo que provoca a destruição da cartilagem articular e leva a uma deformidade da articulação CAMANHO (2001), as causas da artrose podem ser tanto primárias quanto secundárias, as primárias o processo degenerativo, uma de suas causas está correlacionada com o envelhecimento, as causas secundárias, estão correlacionadas a outros fatores, como por exemplo, a obesidade (FRANCO et al., 2009), cargas excessivas e fatores metabólicos e hormonais.

Atualmente não existe cura conhecida para a artrose, os procedimentos modernos buscam atenuar a dor, minimizar a progressão das lesões e melhorar e prevenir limitações

nas deformidades articulares, garantindo uma melhor qualidade de vida ao paciente, logo é de suma importância a realização do diagnóstico nos estágios iniciais.

O diagnóstico em fases iniciais é possível, sendo diagnosticada apenas por uma radiografia, a cartilagem não é visível em imagens de raio-X, com a perda de cartilagem é revelada pelo estreitamento do espaço entre os ossos na articulação, contudo torna-se difícil o diagnóstico em casos iniciais, em casos especiais podem ser utilizados uma ressonância magnética ou utilização de ultrassom (AHO et al., 2017), que garantem uma melhor visualização das perdas de tecidos.

Devido a radiografia ser de menor custo é de maior aplicação para casos de artrose, um outro fator que prejudica o diagnóstico inicial da artrose é a falta de concordância na literatura, sobre a escala de classificação de Kellgren-Lawrence (KL), devido as suas limitações, cujo tal escala proposta por Kellgren-Lawrence (KELLGREN; LAWRENCE, 1957) divide-se entre quatro categorias: Grau 0: Normal, Grau 1: Questionável (Estreitamento duvidoso do espaço articular), Grau 2: Leve, Grau 3: Moderado, Grau 4: Severo.

Algumas críticas a tal modelo proposto por Kellgren-Lawrence, sugere a divergência de descritivos de imagens (KOHN; SASSOON; FERNANDO, 2016), além do fato da ausência de formação de osteófitos não pode ser mensurado no sistema KL (SPECTOR; COOPER, 1993).

Por se tratar de uma doença, cujo não temos uma cura, surge a necessidade de compreender os principais fatores, e desenvolver um modelo de classificação que possa auxiliar na identificação do caso do paciente e prover as melhores medidas, proporcionando um maior bem-estar ao paciente.

Para o desenvolvimento de tal estudo, foi utilizado um conjunto de dados de radiografias com classificações de artrose definidas por Kellgren e Lawrence (1957) disponibilizada pela OAI[5].

5.2.1 Base de Dados

A Base de dados consiste em cerca de 5778 imagens, divididas em 5 categorias, conforme a tabela 7:

A tabela 7, exibe a distribuição de dados e seus respectivos graus, por se tratar de uma base desbalanceada, esse desequilíbrio pode causar problemas durante o processo de classificação do modelo (KUHN; JOHNSON, 2013), com a finalidade de lidar com tal processo utilizaremos técnicas de *Data Augmentation* (GARCEA et al., 2023) nas classes minoritárias, com a finalidade garantir uma equidade de dados de diferentes classes, tais técnicas serão abordadas na seção B, tais dados serão divididos em três conjuntos, o primeiro para treinamento do modelo, o segundo para validação e terceiro para teste do modelo, a figura 5, exibe uma amostra das imagens presentes na base de estudo.

[5] *Knee Osteoarthritis Severity Grading Dataset*: <https://data.mendeley.com/datasets/56rmx5bjcr/1>

Tabela 7 – Distribuição de Dados por Grau de Artrose

Classificação	Volume
Grau 0	2286
Grau 1	1046
Grau 2	1516
Grau 3	757
Grau 4	173
Total	5778

5.2.2 *Pipeline* para a Construção do Modelo

5.2.2.1 Pré-processamento

Com a finalidade de garantir um maior balanceamento entre os dados, foram realizadas operações nas classes minoritárias com a finalidade de aumentar o tamanho e a diversidade presente nessas classes. As operações utilizadas foram:

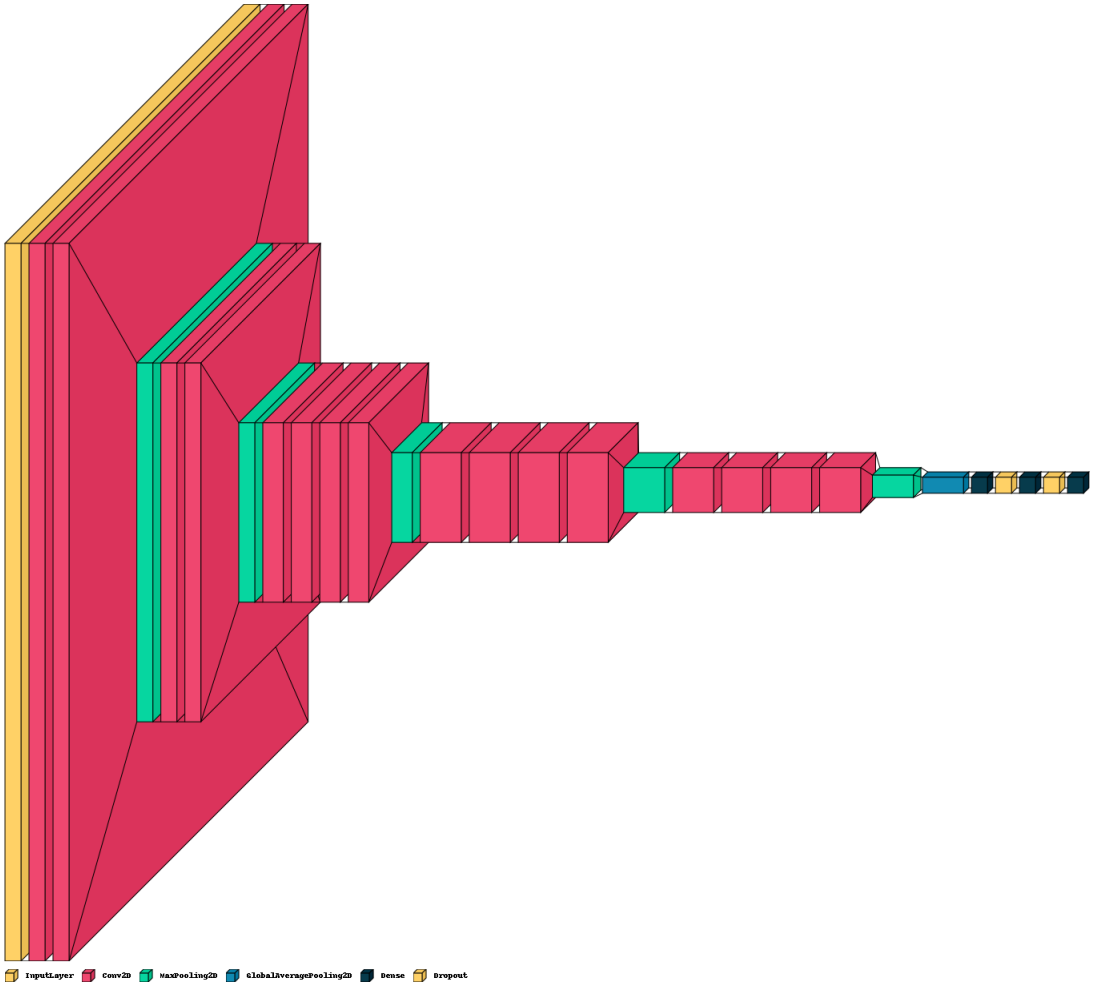
- **Rotação**: Foi realizado uma rotação no conjunto de imagens, 10 graus para a direita ou esquerda, a figura 6 exibe um exemplo de rotação;
- ***width shift range***: Foi utilizado um valor de 0.05 para deslocamento de imagens, esse valor representa a largura total que a imagem pode ser deslocada.
- ***height shift range***: Utilizado um valor de 0.05 para deslocamento de imagens, esse valor representa altura que a imagem pode ser deslocada.
- ***horizontalf lip***: Método que permite fazer o espelhamento da imagem, a figura x exibe um exemplo de tal método.
- **Correção de gamma com ruído**: Método que combina a correção de gamma com aplicação de ruído gaussiano permitindo a maior robustez do modelo.

5.2.2.2 Modelo Proposto

Utilizaremos uma variação do modelo do VGG19, utilizaremos a técnica de *Transfer Learning*, técnica amplamente utilizada em modelos de aprendizado de máquina e visão computacional. Usando o conhecimento já conhecido do modelo, é útil para novas tarefas, especialmente em situações em que temos escassez de dados.

A partir do modelo do VGG 19, Figura 47, as últimas oito camadas do modelo foram congeladas para que não sejam treinadas novamente durante o treinamento do

Figura 47 – Arquitetura VGG19. Fonte: Autoria Própria.



modelo personalizado, adicionamos uma camada de ***GlobalAveragePooling2D*** é usado para adicionar uma camada de ***pooling*** global média.

Ao calcular a média dos recursos em cada mapa de ativação, essa camada reduz a dimensionalidade dos recursos espaciais em um vetor de tamanho fixo. São adicionadas duas camadas densamente conectadas, uma com 128 neurônios e outra com 64, também foi utilizado a função de ***Dropout*** para evitar o ***overfitting*** Após a camada densa inicial, uma camada de saída inicial com taxa de saída de 0.4 é adicionada.

Após a segunda camada densa, é inserida uma segunda camada de *dropout* com uma taxa de 0,2. O *droupout*ajuda a prevenir *overfitting* ao desativar aleatoriamente uma porcentagem das unidades de saída durante o treinamento. Para a camada de saída, foram utilizados 5 neurônios, correspondentes às 5 classes presentes no estudo.

Utilizaremos o otimizador Gradiente Descendente Estocástico (***Stochastic Gradient Descent*** - - SGD), método que busca otimizar a descida do gradiente durante cada busca, uma vez que o vetor de pesos aleatório é escolhido. é utilizado para reduzir a função de perda, ajustando os pesos da rede de acordo com a direção oposta ao gradiente da função de perda.

Tabela 8 – Resultados Experimentais. Fonte: Autoria Própria.

	Acurácia	Precision	Recall	Cohen	F1-Score
Experimento 1	59.00	59.88	62.23	47.78	62.24
Experimento 2	60.59	61.11	64.30	49.23	63.93
Experimento 3	65.70	62.86	66.48	51.41	63.96
Experimento 4	68.30	66.48	68.11	54.47	64.70

O gradiente é calculado usando retropropagação (*backpropagation*), e a taxa de aprendizado e o gradiente são usados para ajustar os pesos, como parâmetros foi utilizado uma taxa de aprendizado de 0.0001, a taxa de aprendizado regula a quantidade de passos que o otimizador faz ao alterar os parâmetros do modelo, uma taxa de aprendizado elevada, resulta em uma convergência maior, contudo pode resultar oscilações globais mínimas, uma taxa de aprendizado menor reduz os passos e a convergência, tornando mais suave e levando mais tempo para atingir o mínimo global.

Momentum, foi utilizado com uma taxa de 0.9, tal métrica busca regular o efeito da inércia ao ajustar os parâmetros, permitindo que o otimizador acumule o gradiente ao longo de várias iterações e continue movendo-se na direção certa, mesmo que a superfície de perda seja irregular evitando mínimos locais indesejados e acelerando a convergência.

5.2.2.3 Experimentos

Para validação das técnicas previamente citadas, iremos realizar 4 experimentos, tais variações de experimentos, buscam acrescentar o número de imagens, nas classes minoritárias, através dos processos de *Data Augmentation*, os experimentos seguem da seguinte forma:

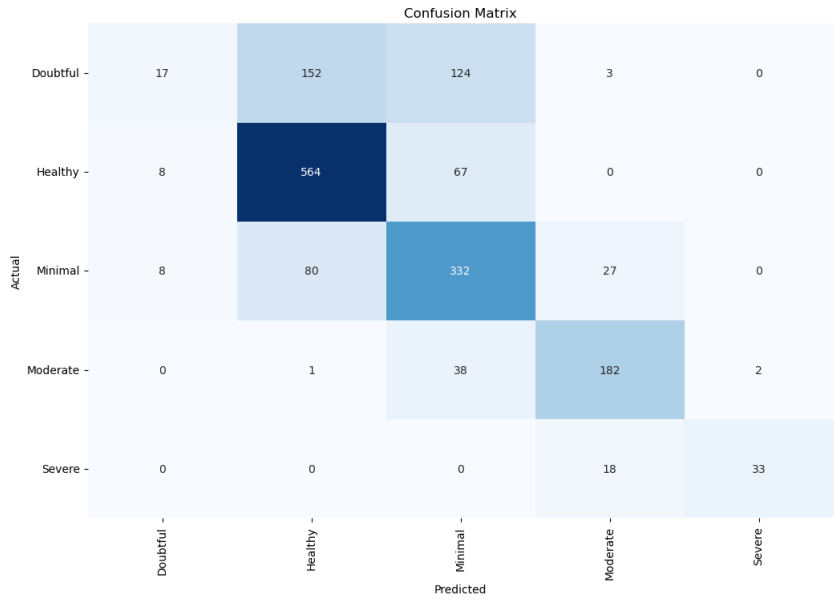
- **Experimento 1**: Base tratada a partir das técnicas de Data Augmentation, com limite de 500 amostras em cada classe;
- **Experimento 2**: Base tratada a partir das técnicas de Data Augmentation, com limite de 1000 amostras em cada classe;
- **Experimento 3**: Base tratada a partir das técnicas de Data Augmentation, com limite de 1500 amostras em cada classe;
- **Experimento 4**: Base tratada a partir das técnicas de Data Augmentation, com limite de 2000 amostras em cada classe, resultando em uma nova base com 10000 amostras.

A partir de tais experimentos é possível entender os impactos do processo de balanceamento de dados e como tais manipulações afetam o modelo.

Como podemos visualizar, na tabela 8, que o Experimento 4, foi o que obteve o maior sucesso entre os experimentos estudados, com uma maior base e uma maior

variação de dados, foi possível que nosso modelo proposto obtendo melhores resultados em comparativo com os modelos propostos por Tiulpin et al. (2018), com uma acurácia de 67.49% e 66.71% para a tarefa de multi-classificação e Górriz et al. (2019), com cerca de 64.30% de acurácia, na qual, ambos se utilizaram de uma combinação de duas bases, com mais de 15 mil imagens apenas no conjunto de treino, enquanto o nossa base de treino para o nosso melhor experimento continha apenas 10000 imagens. A figura 48, exibe a matriz confusão para o experimento 4.

Figura 48 – Matriz Confusão Experimento 4. Fonte: Autoria Própria.



5.2.3 Aplicação e Avaliação dos Resultados

Para simular o comportamento do usuário, será enviada uma imagem de osteoartrite no formato (.png). O objetivo é demonstrar a flexibilidade da plataforma ao lidar com diversos tipos de dados.

Essa etapa, como mencionado anteriormente, é realizada pelo ***Data Type Extractor***, que interpreta dados provenientes de diferentes fontes, a figura 49, exibe o dado amostral que será utilizado para a tarefa de predição.

Por se tratar de um dado de imagem e o modelo treinado ser uma rede neural, logo os filtros determinam ***Data Type Extractor Task Type Extractor***, que será gerado uma explicação local, conforme visualizado na figura 51.Neste estudo de caso, iremos utilizar duas técnicas de geração de explicabilidade, nais quais são o LIME e o GRAD-CAM, ambas são técnicas de explicação visuais, para o domínio de imagens.

Ao selecionar o método de explicabilidade, utilizamos a técnica Grad-CAM para gerar a visualização, figura 52. Essa abordagem permite a criação de um mapa de calor que destaca as regiões com maior relevância para a tarefa de classificação. O mapa de calor

Figura 49 – Dado amostral, osteoartrite nível 4. Fonte: Autoria Própria.

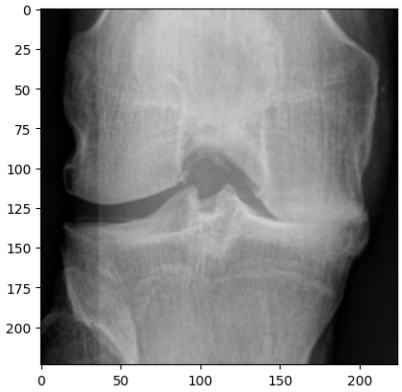


Figura 50 – Página inicial, com leitura de um dado de imagem. Fonte: Autoria Própria.

indica as áreas da imagem que tiveram o maior impacto na decisão do modelo, facilitando a interpretação dos resultados.

De maneira similar, podemos escolher a técnica LIME, que oferece uma abordagem alternativa para explicabilidade Figura 53. No caso de imagens, o LIME funciona segmentando a imagem em *superpixels* e perturbando essas regiões de forma aleatória. Em seguida, o método avalia o impacto de cada *superpixel* na predição do modelo, destacando as áreas que mais influenciam a decisão. Assim, é possível visualizar quais partes da imagem contribuem positivamente ou negativamente para a classificação, facilitando a interpretação e compreensão do comportamento do modelo.

Figura 51 – Técnicas possíveis de explicabilidade visual. Fonte: Autoria Própria.

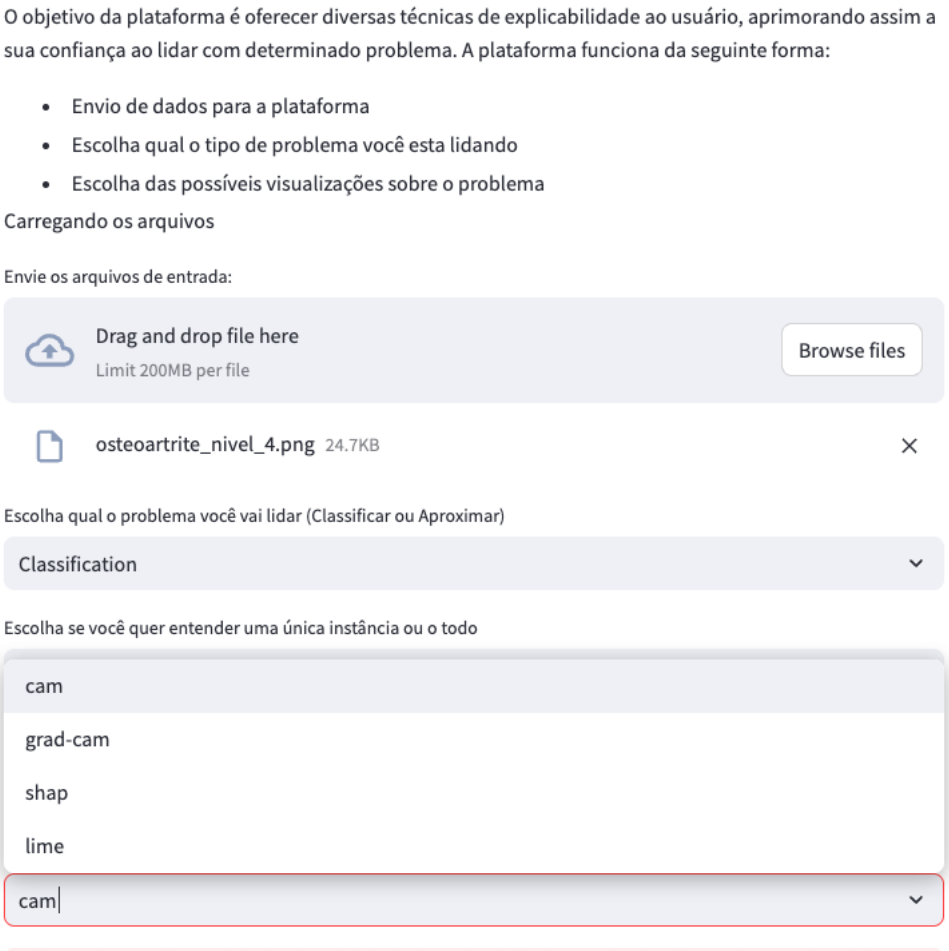
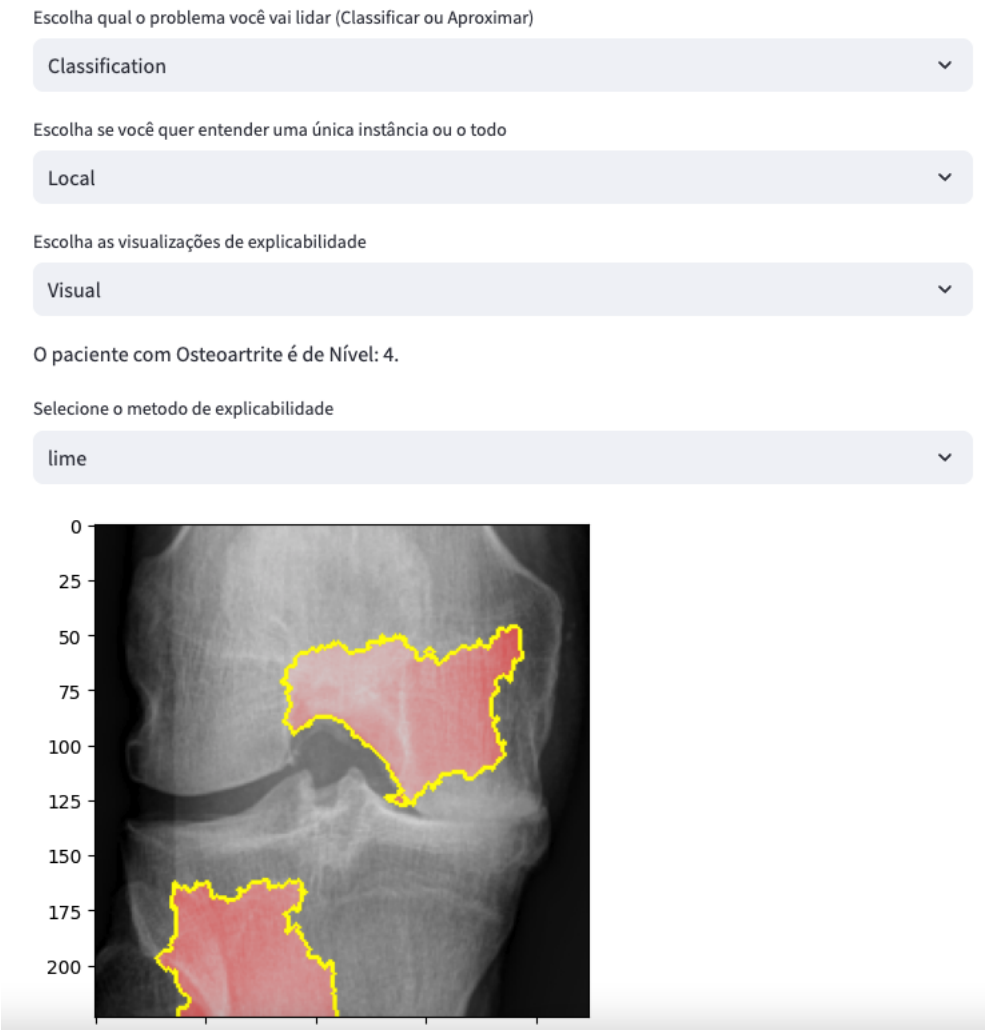


Figura 52 – Visualização explicabilidade Grad-CAM. Fonte: Autoria Própria.

Figura 53 – Visualização explicabilidade LIME. Fonte: Autoria Própria.



6 Conclusões e Trabalhos Futuros

Neste capítulo, apresentamos as conclusões do presente trabalho, discutimos as limitações da pesquisa, destacamos as contribuições e sugerimos direções para trabalhos futuros. A seção de conclusões resume os principais achados e implicações do estudo. As limitações abordam as restrições e desafios enfrentados durante a pesquisa, enquanto as contribuições destacam os avanços e *insights* gerados. Por fim, a seção de trabalhos futuros propõe áreas e questões que podem ser exploradas em pesquisas subsequentes para aprofundar o entendimento e aprimorar as práticas na área de estudo.

6.1 Contribuições

Embora a abordagem AIaaS seja relativamente nova no meio acadêmico, ela já está estabelecida na indústria. Essa prática, que visa a criação e disponibilização de modelos de IA como serviços, tem se mostrado eficaz na implementação de soluções de inteligência artificial em larga escala.

Portanto, espera-se que a arquitetura de referência proposta para AIaaS com ênfase em explicabilidade traga várias contribuições significativas:

1. Elaboração de uma arquitetura de referência fundamentada em AIaaS e explicabilidade voltada para o domínio da saúde, visando aprimorar a compreensão de modelos e processos de tomada de decisão. Essa abordagem busca não apenas promover a transparência dos modelos, mas também a sua interpretabilidade, permitindo que usuários entendam melhor as operações e resultados dos sistemas de inteligência artificial.

2. A arquitetura inovadora oferece uma base sólida para o desenvolvimento de novas aplicações de IA que exigem explicabilidade. Ela pode ser adaptada e expandida para atender a uma variedade de casos de uso em diferentes setores.

3. Fornecendo um registro transparente e passível de auditoria do processo de desenvolvimento do modelo, a arquitetura assegura aos profissionais de saúde maior confiança na aplicação de aprendizado de máquina.

4. A arquitetura proporciona mecanismos para garantir que os modelos de IA atendam aos requisitos regulatórios relacionados à transparência e à explicabilidade. Isso inclui a documentação e a explicação das decisões dos modelos de acordo com as regulamentações vigentes.

6.2 Limitações da Abordagem

Ao longo da realização desta pesquisa, foram identificadas algumas possíveis limitações à arquitetura de referência sugerida:

- Os dados de saúde frequentemente estão armazenados em diversos formatos e sistemas, o que dificulta sua integração e análise;
- A necessidade de fornecer explicações personalizadas e compreensíveis para diferentes partes interessadas, a dificuldade em lidar com a complexidade dos modelos de caixa preta e a necessidade de alinhar os modelos com o contexto e as práticas de trabalho do domínio em que serão utilizados;
- Algumas técnicas de explicabilidade podem não escalar bem com grandes volumes de dados ou modelos complexos, resultando em explicações que são difíceis de gerar ou interpretar em larga escala .
- A implementação do AIaaS pode ser custosa, uma vez que exige investimentos em infraestrutura e ferramentas. Isso pode representar um obstáculo para organizações menores ou aquelas com orçamentos limitados.
- A arquitetura AIaaS pode não suportar ou integrar facilmente ferramentas e técnicas de explicabilidade específicas. Isso pode ocorrer devido à falta de compatibilidade ou suporte para *frameworks* de explicabilidade.
- Fornecer explicações detalhadas pode exigir acesso a dados sensíveis, o que pode levantar questões de privacidade e segurança.

6.3 Conclusões

O presente trabalho teve como objetivo apresentar a arquitetura de H-XAIaaS, uma arquitetura baseada em AIaaS, voltada ao domínio da saúde.

Foi realizada uma descrição detalhada de cada componente da arquitetura proposta juntamente com suas principais características e interfaces. No que se refere à revisão bibliográfica, foi realizada uma revisão narrativa da literatura, com objetivo de examinar trabalhos que propunham abordagens similares à arquitetura proposta.

Embora a definição de AIaaS seja recente, ela demonstra promissora no contexto da criação de modelos de aprendizado de máquina. Além disso, há uma perspectiva de complementaridade com técnicas de explicabilidade, ampliando a transparência no processo decisório do modelo. Isso não se limita apenas à interpretabilidade do próprio modelo, mas sim abre caminho para um entendimento mais profundo de todo o processo, reforçando sua importância no cenário da tomada de decisões.

A partir dos dois estudos de caso, foi possível entender como a arquitetura H-XAIaaS se comporta para conjuntos de dados distintos, permitindo observar a sua adaptabilidade. É importante destacar que as explicações geradas estão intimamente correlacionadas à qualidade do modelo.

Por fim, os resultados deste estudo estabelecem um caminho promissor para o campo da saúde, permitindo que modelos de aprendizado de máquina se transformem em produtos eficazes em vez de experimentos simples. A adoção dessa abordagem juntamente com técnicas de explicabilidade não apenas aumenta a confiabilidade desses modelos, mas também permite que os profissionais de saúde entendam melhor as decisões que são tomadas.

6.4 Trabalhos Futuros

Para aumentar a robustez e a generalização da arquitetura proposta, trabalhos futuros devem explorar a validação em conjuntos de dados de diferentes domínios, além do setor médico, como os setores de marketing, financeiro, etc. Essa diversificação permitirá avaliar o desempenho e a adaptabilidade da arquitetura em contextos variados, proporcionando uma visão mais abrangente de sua aplicabilidade e eficiência.

Adicionalmente, é crucial desenvolver mecanismos que promovam uma interação mais dinâmica e intuitiva com a explicabilidade do modelo. Isso pode incluir a criação de interfaces que facilitem a cooperação homem-máquina, oferecendo explicações interativas que vão além das representações estáticas propostas neste trabalho. Tais avanços não só aprimorariam a experiência do usuário, mas também poderiam aumentar a confiança e a compreensão dos sistemas de inteligência artificial em contextos práticos.

Referências

AHO, O.-M. et al. Subchondral bone histology and grading in osteoarthritis. *PLOS ONE*, Public Library of Science, v. 12, n. 3, p. 1–16, 03 2017. Disponível em: <https://doi.org/10.1371/journal.pone.0173726>.

ANGELOV, P. et al. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 11, 07 2021.

ARAUJO, K. L. R. d. et al. Fatores associados à síndrome respiratória aguda grave em uma região central do brasil. *Ciência Saúde Coletiva*, ABRASCO - Associação Brasileira de Saúde Coletiva, v. 25, p. 4121–4130, Oct 2020. ISSN 1413-8123. Disponível em: <https://doi.org/10.1590/1413-812320202510.2.26802020>.

ASSADI, A. et al. An integration engineering framework for machine learning in healthcare. *Frontiers in Digital Health*, v. 4, 2022. ISSN 2673-253X. Disponível em: <https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2022.932411>.

BAGGIO JUSSARA A. OLIVEIRA; EXEL, A. L. C. A. C. d. N. M. V. Síndrome respiratória aguda grave (srag) causada por covid-19: Um fator regional. *Arq. Bras. Cardiol., v. 117, n. 5, p. 976-977*, 2021.

BARATCHI M., W. C. L. S. e. a. Automated machine learning: past, present and future. *Artif Intell Rev 57, 122*, 2024.

BARRETO, R. G. et al. Clinical decision support based on owl queries in a knowledge-as-a-service architecture. In: _____. [S.l.: s.n.], 2018. p. 226–238. ISBN 978-3-319-99905-0.

BEEMAN, D. *Multi-layer perceptrons and back propagation*. 2001. <https://www.cs.cmu.edu/afs/club/user/cmccabe/ecee.colorado.edu/~ecen4831/lectures/NNet3.html>.

BELLE, V.; PAPANTONIS, I. Principles and practice of explainable machine learning. *Frontiers in Big Data*, v. 4, 2021. ISSN 2624-909X. Disponível em: <https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2021.688969>.

BHARATI, S.; MONDAL, M. R. H.; PODDER, P. A review on explainable artificial intelligence for healthcare: Why, how, and when? *IEEE Transactions on Artificial Intelligence*, Institute of Electrical and Electronics Engineers (IEEE), v. 5, n. 4, p. 1429–1442, abr. 2024. ISSN 2691-4581. Disponível em: <http://dx.doi.org/10.1109/TAI.2023.3266418>.

BHATTACHARYA, S. et al. Artificial intelligence enabled healthcare: A hype, hope or harm. *Journal of Family Medicine and Primary Care*, v. 8, p. 3461, 11 2019.

BIRAN, O.; COTTON, C. V. Explanation and justification in machine learning : A survey or. In: . [s.n.], 2017. Disponível em: <https://api.semanticscholar.org/CorpusID:3911355>.

BORNIA C., B. A. R. M. *Estatística para Cursos de Engenharia e Informática*. [S.l.: s.n.], 2010.

BRAGA, A. d. P.; LUDERMIR, T. B.; CARVALHO, A. C. P. d. L. F. d. *Redes neurais artificiais: teoria e aplicações*. [S.l.]: LTC, 2000.

BRASIL, M. da S. *Nota Técnica nº 31/2022-CGPNI/DEIDT/SVS/MS. Informações técnicas e recomendações sobre a vigilância epidemiológica da Influenza no Brasil.* Brasília, DF, 2022. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/covid-19/notas-tecnicas/2022/nota-tecnica-no-31-2022-cgpni-deidt-svs-ms.pdf>.

BREIMAN, L. Random forests. Springer, v. 45, n. 1, p. 5–32, 2001.

BROWN, S. *Machine learning, explained | MIT Sloan — mitsloan.mit.edu*. 2021. <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>. [Acessado 08-08-2024].

BRUNESE, L. et al. Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays. *Computer Methods and Programs in Biomedicine*, v. 196, p. 105608, 2020. ISSN 0169-2607. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169260720314413>.

BöHLE, M. et al. Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification. *Frontiers in Aging Neuroscience*, Frontiers Media SA, v. 11, jul. 2019. ISSN 1663-4365. Disponível em: <http://dx.doi.org/10.3389/fnagi.2019.00194>.

CALISKAN, A.; BRYSON, J.; NARAYANAN, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, v. 356, p. 183–186, 04 2017.

CAMANHO, G. *Revista Brasileira de Ortopedia - Tratamento da osteoartrose do joelho*. Revista Brasileira de Ortopedia - Tratamento da osteoartrose do joelho, 2001. Disponível em: <https://rbo.org.br/detalhes/107/pt-BR/tratamento-da-osteoartrose-do-joelho>.

CAMANHO, G. L.; IMAMURA, M.; ARENDT-NIELSEN, L. Gênese da dor na artrose. *Revista Brasileira de Ortopedia*, Sociedade Brasileira de Ortopedia e Traumatologia, v. 46, n. 1, p. 14–17, 2011. ISSN 0102-3616. Disponível em: <https://doi.org/10.1590/S0102-36162011000100002>.

CARUANA, R. et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015. Disponível em: <https://api.semanticscholar.org/CorpusID:14190268>.

CASTRO, L. de; FERRARI, D. *Introdução a mineração de dados*. Editora Saraiva, 2017. ISBN 9788547200992. Disponível em: <https://books.google.com.br/books?id=SSlrDwAAQBAJ>.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016. Disponível em: <http://arxiv.org/abs/1603.02754>.

CLARKE, E. *Back Propagation in Neural Network: Machine Learning Algorithm*. 2024. <https://www.guru99.com/backpropogation-neural-network.html$>.

CLEMENTS, P. C. Software architecture in practice. *Diss. Software Engineering Institute*, 2002.

COROAMă, L.; GROZA, A. Evaluation metrics in explainable artificial intelligence (xai). In: _____. [S.l.: s.n.], 2022. p. 401–413. ISBN 978-3-031-20318-3.

CROSS, M. et al. The global burden of hip and knee osteoarthritis: Estimates from the global burden of disease 2010 study. *Annals of the rheumatic diseases*, v. 73, 02 2014.

DAS, S. et al. Taxonomy and survey of interpretable machine learning method. In: . [S.l.: s.n.], 2020.

DASGUPTA, S.; FROST, N.; MOSHKOVITZ, M. *Framework for Evaluating Faithfulness of Local Explanations*. 2022. Disponível em: <https://arxiv.org/abs/2202.00734>.

DINEVSKI, D. et al. Clinical decision support systems. In: _____. [S.l.: s.n.], 2011. ISBN 978-953-307-354-5.

DIPROSE, W. et al. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association*, v. 27, 02 2020.

DOSHI-VELEZ, F.; KIM, B. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. Disponível em: <https://arxiv.org/abs/1702.08608>.

ELSHAWI, R.; SAKR, S. *Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service*. 2017. Disponível em: <https://arxiv.org/abs/1709.07493>.

EUROPEIA, U. Lei nº 1689, de 12 de julho de 2024. *Official Journal of the European Union*, União Europeia, 2024. Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689>.

FAYYAD, U. Knowledge discovery in databases: An overview. In: _____. *Relational Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. p. 28–47. ISBN 978-3-662-04599-2. Disponível em: <https://doi.org/10.1007/978-3-662-04599-2_2>.

FIELDING R.T; RICHARD, N. *Architectural styles and the design of network-based software architectures*. Tese (Doutorado) — University of California, Irvine, 2000.

FIELDING R.T; RICHARD, N. *Architectural styles and the design of network-based software architectures.* Tese (Doutorado) — University of California, Irvine., 2000.

FONTANA, A. *The AI-First Company: How to Compete and Win with Artificial Intelligence*. 2021.

FRANCO, L. R. et al. Influência da idade e da obesidade no diagnóstico sugestivo de artrose de joelho. *ConScientiae Saúde*, v. 8, n. 1, p. 41–46, maio 2009. Disponível em: <https://periodicos.uninove.br/saude/article/view/1506>.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. In: *Proceedings of the Second European Conference on Computational Learning Theory*. Berlin, Heidelberg: Springer-Verlag, 1995. (EuroCOLT '95), p. 23–37. ISBN 3540591192.

GARCEA, F. et al. Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, v. 152, p. 106391, 2023. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S001048252201099X>.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. *Machine Learning*, v. 63, p. 3–42, 04 2006.

GODDARD, K.; ROUDSARI, A.; WYATT, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, v. 19, n. 1, p. 121–127, 06 2011. ISSN 1067-5027. Disponível em: <https://doi.org/10.1136/amiajnl-2011-000089>.

GREENES, R. *Clinical Decision Support: The Road Ahead*. Elsevier Science, 2011. ISBN 9780080467696. Disponível em: <https://books.google.com.br/books?id=_f4u1I-c6PIC>.

GROSSMAN, R. L.; HORNICK, M. F.; MEYER, G. Data mining standards initiatives. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 45, n. 8, p. 59–61, aug 2002. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/545151.545180>.

GUNNING D.; VORM, E. W. J.; TUREK, M. Darpa 's explainable ai (xai) program: A retrospective. *Applied AI Letters*, v. 2, 12 2021.

GóRRIZ, M. et al. Assessing knee oa severity with cnn attention-based end-to-end architectures. In: . [S.l.: s.n.], 2019.

HUANG, G.-B.; ZHU, Q.-Y.; SIEW, C.-K. Extreme learning machine: Theory and applications. *Neurocomputing*, v. 70, n. 1, p. 489–501, 2006. ISSN 0925-2312. Neural Networks. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231206000385>.

HUANG, I.; LIM, M.; PRANATA, R. Diabetes mellitus is associated with increased mortality and severity of disease in covid-19 pneumonia – a systematic review, meta-analysis, and meta-regression. *Diabetes Metabolic Syndrome: Clinical Research Reviews*, v. 14, 04 2020.

JUNG, J. et al. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon*, v. 9, p. e16110, 05 2023.

KADIR, B. A.; BROBERG, O. Human-centered design of work systems in the transition to industry 4.0. *Applied Ergonomics*, v. 92, p. 103334, 2021. ISSN 0003-6870. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0003687020302829>.

KELLGREN, J. H.; LAWRENCE, J. S. Radiological assessment of osteo-arthrosis. *Annals of the Rheumatic Diseases*, v. 16, p. 494 – 502, 1957. Disponível em: <https://api.semanticscholar.org/CorpusID:45581335>.

KHATTAK, F. K. et al. *MLHOps: Machine Learning for Healthcare Operations*. 2023. Disponível em: <https://arxiv.org/abs/2305.02474>.

KOHN, M.; SASSOON, A.; FERNANDO, N. Classifications in brief: Kellgren-lawrence classification of osteoarthritis. *Clinical Orthopaedics and Related Research*, v. 474, 02 2016.

KOZIARSKI, M. *Subamostragem baseada em radial para classificação de dados desequilibrados*. [S.l.]: Reconhecimento de padrões, v. 102, n. 3, 2020.

KUHN, M.; JOHNSON, K. *Applied Predictive Modeling,*. [S.l.]: Springer, 2013.

LANGER, M. et al. What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, Elsevier BV, v. 296, p. 103473, jul. 2021. ISSN 0004-3702. Disponível em: <http://dx.doi.org/10.1016/j.artint.2021.103473>.

LEE, J.; NISHIKAWA, R. M. Detecting mammographically-occult cancer in women with dense breasts using deep convolutional neural network and radon cumulative distribution transform. In: MORI, K.; HAHN, H. K. (Ed.). *Medical Imaging 2019: Computer-Aided Diagnosis*. SPIE, 2019. v. 10950, p. 1095003. Disponível em: <https://doi.org/10.1117/12.2512446>.

LEMNARU, C.; POTOLEA, R. Imbalanced classification problems: Systematic study, issues and best practices. In: ZHANG, R. et al. (Ed.). *Enterprise Information Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 35–50. ISBN 978-3-642-29958-2.

LEWICKI, K. et al. Out of context: Investigating the bias and fairness concerns of "artificial intelligence as a service". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, 2023. (CHI '23, v. 165), p. 1–17. Disponível em: <http://dx.doi.org/10.1145/3544548.3581463>.

LINS, S. et al. Artificial intelligence as a service. *Business Information Systems Engineering*, v. 63, 08 2021.

LIPTON, Z. C. *The Mythos of Model Interpretability*. 2017. Disponível em: <https://arxiv.org/abs/1606.03490>.

LU, Q. et al. Towards a roadmap on software engineering for responsible ai. In: *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*. New York, NY, USA: Association for Computing Machinery, 2022. (CAIN '22), p. 101–112. ISBN 9781450392754. Disponível em: <https://doi.org/10.1145/3522664.3528607>.

LUSS, R. et al. *Generating Contrastive Explanations with Monotonic Attribute Functions*. 2019.

LóPEZ, V. et al. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, v. 250, p. 113–141, 2013. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025513005124>.

MARKOV, I. L. et al. Looper: An end-to-end ml platform for product decisions. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2022. (KDD '22), p. 3513–3523. ISBN 9781450393850. Disponível em: <https://doi.org/10.1145/3534678.3539059>.

MARR, D. Artificial intelligence—a personal view. *Artif. Intell.*, Elsevier Science Publishers Ltd., GBR, v. 9, n. 1, p. 37–48, aug 1977. ISSN 0004-3702. Disponível em: <https://doi.org/10.1016/0004-3702(77)90013-3>.

MASÍS, S. *Interpretable Machine Learning with Python - Second Edition*. [S.l.]: Packt Publishing, 2021.

MCDERMID, J. et al. Artificial intelligence explainability: the technical and ethical dimensions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 379, p. 20200363, 08 2021.

MCGOVERN, J. et al. *Enterprise Service Oriented Architectures: Concepts, Challenges, Recommendations*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 140203704X.

MERJULAH, R.; CHANDRA, J. Chapter 10 - classification of myocardial ischemia in delayed contrast enhancement using machine learning. In: HEMANTH, D. J.; GUPTA, D.; Emilia Balas, V. (Ed.). *Intelligent Data Analysis for Biomedical Applications*. Academic Press, 2019, (Intelligent Data-Centric Systems). p. 209–235. ISBN 978-0-12-815553-0. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128155530000112>.

MERTES, S. et al. Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in Artificial Intelligence*, Frontiers Media SA, v. 5, abr. 2022. ISSN 2624-8212. Disponível em: <http://dx.doi.org/10.3389/frai.2022.825565>.

MESKE, C. et al. Explainable artificial intelligence: Objectives, stakeholders and future research opportunities. *Information Systems Management*, 12 2020.

METTA, C. et al. Towards transparent healthcare: Advancing local explanation methods in explainable artificial intelligence. *Bioengineering*, v. 11, n. 4, 2024. ISSN 2306-5354. Disponível em: <https://www.mdpi.com/2306-5354/11/4/369>.

MIRANDA, E.; ARYUNI, M.; IRWANSYAH, E. *A Survey of Medical Image Classification Techniques*. 2016.

MITCHELL, T. *Machine Learning*. McGraw-Hill, 1997. (McGraw-Hill International Editions). ISBN 9780071154673. Disponível em: <https://books.google.com.br/books?id=EoYBngEACAAJ>.

MOLNAR, C. *Interpretable Machine Learning*: A guide for making black box models explainable. 2. ed. [s.n.], 2022. Disponível em: <https://christophm.github.io/interpretable-ml-book>.

MONTALVÃO, E. A. *A sobrevida dos casos Hospitalizados de Síndrome Respiratória Aguda Grave (SRAG) por COVID-19 no município do Rio de Janeiro, Brasil, 2020 e 2021.* Tese (Doutorado) — Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, 2022.

MONTENEGRO, T.; LINO, N. Xh-kaas (explainable health-knowledge as a service). In: . [S.l.: s.n.], 2024. p. 309–314.

MOXON, B. Defining data mining. *DBMS*, Miller Freeman, Inc., USA, v. 9, n. 9, p. S11–14, aug 1996. ISSN 1041-5173.

MUKHTOROV, D. et al. Endoscopic image classification based on explainable deep learning. *Sensors*, v. 23, n. 6, 2023. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/23/6/3176>.

NAUTA, M. et al. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 55, n. 13s, jul 2023. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3583558>.

NGUYEN, A. phi; MARTíNEZ, M. R. *On quantitative aspects of model interpretability*. 2020. Disponível em: <https://arxiv.org/abs/2007.07584>.

OKAY, F. Y.; YıLDıRıM, M.; ÖZDEMIR, S. Interpretable machine learning: A case study of healthcare. In: *2021 International Symposium on Networks, Computers and Communications (ISNCC)*. [S.l.: s.n.], 2021. p. 1–6.

O'MAHONY, C. et al. A novel clinical risk prediction model for sudden cardiac death in hypertrophic cardiomyopathy (HCM Risk-SCD). *European Heart Journal*, v. 35, n. 30, p. 2010–2020, 10 2013. ISSN 0195-668X. Disponível em: <https://doi.org/10.1093/eurheartj/eht439>.

O'SHEA, K.; NASH, R. *An Introduction to Convolutional Neural Networks*. 2015. Disponível em: <https://arxiv.org/abs/1511.08458>.

PHILIPP, R. et al. Machine learning as a service: Challenges in research and applications. In: *Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications & Services*. New York, NY, USA: Association for Computing Machinery, 2021. (iiWAS '20), p. 396–406. ISBN 9781450389228. Disponível em: <https://doi.org/10.1145/3428757.3429152>.

PONCE, H.; MARTINEZ-VILLASEñOR, M. de L. Interpretability of artificial hydrocarbon networks for breast cancer classification. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. [S.l.: s.n.], 2017. p. 3535–3542.

QUONIAM, L. e. a. Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o brasil. *Ciência da Informação, Brasília, v. 30, n. 2, p. 20-28, maio/ago.*, 2001.

RANZANI, O. T. e. a. Characterisation of the first 250000 hospital admissions for covid-19 in brazil: a retrospective analysis of nationwide data. *The Lancet Respiratory Medicine,*, 2021.

RASHED-AL-MAHFUZ, M. et al. Clinically applicable machine learning approaches to identify attributes of chronic kidney disease (ckd) for use in low-cost diagnostic screening. *IEEE Journal of Translational Engineering in Health and Medicine*, v. 9, p. 1–11, 2021.

RECOMMENDATIONS for the medical management of osteoarthritis of the hip and knee: 2000 update. [s.n.], 2000. v. 43. 1905-1915 p. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/1529-0131%28200009%2943%3A9%3C1905%3A%3AAID-ANR1%3E3.0.CO%3B2-P>.

REUMATOLOGIA, S. B. de. Osteoartrite (artrose) - sociedade brasileira de reumatologia. *Sociedade Brasileira de Reumatologia*, 2017. Disponível em: <https://www.reumatologia.org.br/doencas-reumaticas/osteoartrite-artrose/>.

RIBEIRO, M.; GROLINGER, K.; CAPRETZ, M. Mlaas: Machine learning as a service. In: . [S.l.: s.n.], 2015.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 2016. Disponível em: <https://arxiv.org/abs/1602.04938>.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 32, n. 1, Apr. 2018. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/11491>.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3. ed. [S.l.]: Prentice Hall, 2010.

SEKI, T.; KAWAZOE, Y.; OHE, K. Machine learning-based prediction of in-hospital mortality using admission laboratory data: A retrospective, single-site study using electronic health record data. *PLOS ONE*, v. 16, p. e0246640, 02 2021.

SELVARAJU, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, Springer Science and Business Media LLC, v. 128, n. 2, p. 336–359, out. 2019. ISSN 1573-1405. Disponível em: <http://dx.doi.org/10.1007/s11263-019-01228-7>.

SETTOUTI, N.; CHIKH, M.; SAIDI, M. Generating fuzzy rules for constructing interpretable classifier of diabetes disease. *Australasian physical engineering sciences in medicine / supported by the Australasian College of Physical Scientists in Medicine and the Australasian Association of Physical Sciences in Medicine*, v. 35, p. 257–70, 08 2012.

SHEN, D.; WU, G.; SUK, H.-I. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, Annual Reviews, v. 19, n. Volume 19, 2017, p. 221–248, 2017. ISSN 1545-4274. Disponível em: <https://www.annualreviews.org/content/journals/10.1146/annurev-bioeng-071516-044442>.

SIMONYAN, K.; ZISSERMAN, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. Disponível em: <https://arxiv.org/abs/1409.1556>.

SIWICKI, B. *86% of healthcare companies use some form of AI*. 2017. <https://www.healthcareitnews.com/news/86-healthcare-companies-use-some-form-ai>. [Acessado 08-01-2024].

SPACKMAN, K. Signal detection theory: Valuable tools for evaluating inductive learning. *In: Proc. Sixth Internat. Workshop on Machine Learning.*, p. 160–163., 1989.

SPECTOR, T.; COOPER, C. Radiographic assessment of osteoarthritis in population studies: whither kellgren and lawrence? *Osteoarthritis and Cartilage*, Elsevier BV, v. 1, n. 4, p. 203–206, 1993. ISSN 10634584. Disponível em: <https://cir.nii.ac.jp/crid/1361418520433466368>.

SPEITH, T. A review of taxonomies of explainable artificial intelligence (xai) methods. In: . [S.l.: s.n.], 2022. p. 2239–2250.

SWETS, J.; DAWES, R.; MONAHAN, J. Better decisions through science. *Scientific American*, v. 283, p. 82–7, 11 2000.

TIULPIN, A. et al. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Scientific Reports*, v. 8, 01 2018.

TJOA, E.; GUAN, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, PP, 09 2020.

TORRI, V. *Textual explanations for intuitive machine learning*. Dissertação (Mestrado) — Scuola di ingegneria industriale e dell'informazione, 2021.

TU, R.-H. et al. A scoring system to predict the risk of organ/space surgical site infections after laparoscopic gastrectomy for gastric cancer based on a large-scale retrospective study. *Surgical Endoscopy*, v. 30, n. 7, p. 3026–3034, 2016.

van der Velden, B. H. et al. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, v. 79, p. 102470, 2022. ISSN 1361-8415. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1361841522001177>.

VILONE, G.; LONGO, L. *Explainable Artificial Intelligence: a Systematic Review*. 2020. Disponível em: <https://arxiv.org/abs/2006.00093>.

WACHTER, S.; MITTELSTADT, B.; RUSSELL, C. *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*. 2018. Disponível em: <https://arxiv.org/abs/1711.00399>.

WANG, Z. et al. Xaiport: A service framework for the early adoption of xai in ai model development. In: *Proceedings of the 2024 ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results*. New York, NY, USA: Association for Computing Machinery, 2024. (ICSE-NIER'24), p. 67–71. ISBN 9798400705007. Disponível em: <https://doi.org/10.1145/3639476.3639759>.

XU, S.; ZHANG, W. Knowledge as a service and knowledge breaching. In: *2005 IEEE International Conference on Services Computing (SCC'05) Vol-1*. [S.l.: s.n.], 2005. v. 1, p. 87–94 vol.1.

YAO, Y. et al. Complexity vs. performance: empirical analysis of machine learning as a service. In: *Proceedings of the 2017 Internet Measurement Conference*. New York, NY, USA: Association for Computing Machinery, 2017. (IMC '17), p. 384–397. ISBN 9781450351188. Disponível em: <https://doi.org/10.1145/3131365.3131372>.

ZEINELDIN R.A., K. M. E. Z. e. a. Explainability of deep neural networks for mri analysis of brain tumors. *Int J CARS 17*, 2022.

APÊNDICE A – Dado SRAG HXAI-KaaS

Tabela 9 – Dados Modelo SRAG

VARIÁVEL
CS_SEXO
faixa_etaria
CS_GESTANT
NOSOCOMIAL
FEBRE
GARGANTA
DISPNEIA
DESC_RESP
SATURACAO
DIARREIA
VOMITO
DOR_ABD
FADIGA
PERD_OLFT
PERD_PALA
PUERPERA
CARDIOPATI
HEMATOLOGI
SIND_DOWN
HEPATICA
ASMA
DIABETES
NEUROLOGIC
PNEUMOPATI
IMUNODEPRE
RENAL
OBESIDADE
HOSPITAL
VACINA_COV
VACINA
UTI
SUPORT_VEN
RAIOX_RES
TOMO_RES
ANTIVIRAL
ANTIVIRAL
ANTIVIRAL
EVOLUCAO